Analysis & Design of Data Farming Algorithm For Cardiac Patient Data

Mohd.Shahnawaz
Dept of Computer Science INFINITY
MGMT & ENGG COLLEGE, Sagar, M.P.
India
shahnawaznbd@gmail.com

Kanak Saxena
Dept of Computer Application
SAMRAT ASHOK TECH INSTITUTE,
Vidisha, M.P. India
kanak.saxena@gmail.com

Hari Mohan Pandey
Dept of Computer Science
Middlesex University
London, U.K.
profharimohanpandey@gmail.com

ABSTRACT— Data farming is a process to grow data by applying various statistical, predictions, machine learning and data mining approach on the available data. As data collection cost is high so many times data mining projects use 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 data set and sufficient for extraction of knowledge but sometimes the data set might be narrow and insufficient to extract meaningful knowledge or the data may not even exist. Mining from wide datasets has received wide attention in the available literature. Many models and algorithms for data reduction & feature selection have been developed for wide datasets. Determining or extracting knowledge from a narrow data set (partial availability of data) or in the absence of an existing data set has not been sufficiently addressed in the literature. In this paper we propose an algorithm for data farming, which farm sufficient data from the available little seed data. Classification accuracy of J48 classification for farmed data is achieved better than classification results for the seed data, which proves that the proposed data farming algorithm is effective.

Keywords— Interactive data exploration and discovery, Methodologies and Tools, Data Farming, J48 Classification, Cardiac Patient data, Missing value estimation.

I. INTRODUCTION

Data farming is started from the project alberts [2] [3].Data farming (fertilization, cultivation, plantation, harvesting) [20] is the process of growing data, in the methodology of data farming, large amount of data are generated through simulation of several configurations from large parameter space and then analyzed for patterns [1]. 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 is implemented on MATLAB and farmed datasets are verified for the classification accuracy on the weka. We used J48 classification; it is an open source Java implementation of the C4.5 algorithm in the weka data mining tool. C4.5 builds decision trees from a set of training data in the same way as ID3 using the concept of information entropy. This paper is organized in 5 sections; section 1 Introduction, Sections 2 describes the dataset used in the research, Section 3 proposed methodology and section 4 describes the outcomes as result analysis and finally section 5 concludes the entire paper.

In this paper the nomenclature for naming the farmed dataset is a combination of three factors involved in farming process:

Syntax: farmed_thresold_seedtuples_farmedtuples

Example: farmed_10_100_5k, means 5000 tuples are farmed from the 100 number of seed tuples on the error threshold value 10.

II. CARDIAC DATASET

In this research, we used medical domain data [8]. It is a cardiac patient data having 20 attribute and 558 instances. Descriptions of the attributes are given in the table I. Dose attribute contains the amount of the dose of dobutamine given to the patient in the past. We had only 558 instances in the original dataset, we took randomly 50 and 100 instances to prepared the sample data sample data_50 & sample data_100 respectively.

TABLE I
SEED DATASET ATTRIBUTE

S.No.	Attribute	Particular
1	bhr	BASAL HEART RATE
2	basebp	BASAL BLOOD PRESSURE
3	basedp	BASAL DOUBLE PRODUCT (= BHR X BASEBP)
4	pkhr	PEAK HEART RATE
5	sbp	SYSTOLIC BLOOD PRESSURE
6	dp	DOUBLE PRODUCT (= PKHR X SBP)
7	dose	DOSE OF DOBUTAMINE GIVEN
8	maxhr	MAXIMUM HEART RATE
9	%mphr(b)	% OF MAXIMUM PREDICTED HEART RATE ACHIEVED BY PATIENT
10	mbp	MAXIMUM BLOOD PRESSURE
11	dpmaxdo	DOUBLE PRODUCT ON MAXIMUM DOBUTAMINE DOSE
12	dobdose	DOBUTAMINE DOSE AT WHICH MAXIMUM DOUBLE PRODUCT OCCURED
13	byear	YEAR OF BIRTH
14	age	PATIENT'S AGE
15	gender	PATIENT'S GENDER (MALE = 0)
16	baseEF	BASELINE CARDIAC EJECTION FRACTION (A MEASURE OF THE HEART'S PUMPING EFFICIENCY)
17	dobEF	EJECTION FRACTION ON DOBUTAMINE
18	phat	VALUE OF PHAT
19	deltaEF	DIFFRENCE OF EJECTION FRACTION
20	newpkmphr	NEW PREDICTED HEART RATE ACHIVED BY PATIENT

III. PROPOSED METHODOLOGY

In this paper we proposed a data farming algorithm to grow data from seed dataset. We have a little input seed dataset and but, we need a lot of data for mining purpose. Proposed algorithm generates data with preserving the range of the input seed data. Proposed data farming methodology completes in these steps.

1. Load the input seed data (m tuple and n attribute)
2. Filling of missing values (if any)
3. Predicting some attribute (if any required)
4. Data farming & farmed data repository

In the step 1 we load input seed data to the model, then in step 2 if input seed data have some missing values. These missing values have to be fill by applying appropriate missing data estimation methods [17]. After that in step 3, we predict some attribute to refine the quality of the seed data i.e. reduce the error between actual and predicted values of some attribute by applying regression [18]. Now in step 4, we use this refined dataset to farm more dataset with the algorithm–I.

In this paper we assume that step 2 & step 3 is already done & input seed is complete and satisfactory to perform data farming, hence in this paper concern only step 4. Pseudo code for the proposed data farming algorithm is given below.

Algorithm-1. Data_farming (seed_dataset, k, error_thresold)

//seed_dataset, it contain seed data in n attribute (a_1 , a_2 , a_3 , … 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 contain the farmed data set of each iteration

```
{
    Farmed_data[k][n];
      for i = 1 to n
      {
          Li= Minimum of column i in seed_data;
          Mi= Maximum of column i in seed_data;
          diffi= Li - Mi ;
          lbi = Li – (diffi* error_thresold/100);
          ubi = Li + (diffi* error_thresold/100);
      }
      for i=1 to k
      {
         for j=1 to n
        {
        farm_data (i,j) = randomly generate the data item
        with bounded range [lbi , ubi] for column j;
        }
     return farmed_data;
}
```

The proposed algorithm is implemented on graphical user interface of MATLAB 7.0. Implemented Model takes seed dataset as .CSV (Comma Separated Values) file format and error_theresold rate as input. And, it stores the farmed dataset also in csv file. Running screen shot of the proposed algorithm is given in Figure 2.

IV. RESULTS AND ANALYSIS

In Table II, we enumerate the various experiments of farming data on different combination of threshold values (2, 5, 10), number of seed instances (50,100) & number of farmed data instances (500, 1k, 2k, 5k, 10k). Seed data used in this paper is related to the cardiac patent. This seed data have 20 attribute as given in Table I. We have performed total 30 numbers of experiments to analyze the proposed algorithm. In this table we give the time required in each experiment & save the farmed data with .csv file name as naming convention described.

TABLE II
DATA FARMING RESULT WITH TIME

Error Thereso ld	No. of Seed Tuple	No. of Farmed Tuple	Time	farmed Dataset
2	50	500	1.094	farmed_2_50_500
2	50	1000	2.109	farmed_2_50_1K
2	50	2000	4.266	farmed_2_50_2K
2	50	5000	12.172	farmed_2_50_5K
2	50	10000	31.328	farmed_2_50_10K
2	100	500	1.11	farmed_2_100_500
2	100	1000	2.172	farmed_2_100_1K
2	100	2000	4.36	farmed_2_100_2K
2	100	5000	12.579	farmed_2_100_5K
2	100	10000	31.922	farmed_2_100_10K
5	50	500	1.109	farmed_5_50_500
5	50	1000	2.156	farmed_5_50_1K
5	50	2000	4.266	farmed_5_50_2K
5	50	5000	12.313	farmed_5_50_5K
5	50	10000	31.687	farmed_5_50_10K
5	100	500	1.125	farmed_5_100_500
5	100	1000	2.172	farmed_5_100_1K
5	100	2000	4.375	farmed_5_100_2K
5	100	5000	12.422	farmed_5_100_5K
5	100	10000	31.938	farmed_5_100_10K
10	50	500	1.125	farmed_10_50_500
10	50	1000	2.172	farmed_10_50_1K
10	50	2000	4.406	farmed_10_50_2K
10	50	5000	12.359	farmed_10_50_5K
10	50	10000	31.328	farmed_10_50_10K
10	100	500	1.109	farmed_10_100_500
10	100	1000	2.171	farmed_10_100_1K
10	100	2000	4.406	farmed_10_100_2K
10	100	5000	12.453	farmed_10_100_5K
10	100	10000	32.297	farmed_10_100_10K

Analysis of the proposed algorithm and factor affecting the performance of the proposed algorithm may be described in points.

- We can observe from the Table II that time required to farm a dataset is highly dependent on the factor that how much instances to be farmed (number of farmed instances). As more instances to be farmed as much time is required.
- 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 is also increases.

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

To check the quality of the farmed datasets we performed classification and compared the classification accuracy among the original dataset, sample datasets and farmed datasets. Here, we used J48 classification in weka. Table III enumerates the result of the classification experiments. TP rate – true positive rate has increased from original dataset to farmed datasets. We have compared original dataset of cardiac patient from medical domain having 20 attribute & 558 instances, a portion i.e. 50 instance as Sample dataset (sample data_50) & 100 instances as sample dataset (sample data_100). We can see the results; TP rate for the farmed datasets has increased compare to the original dataset & sample datasets.

Figure 1 shows the graphical view of the variation in TP rate for the J48 classification, classification is based on attribute "dose". Figure 3 shows the running screen shot of the weka tool while performing classification.

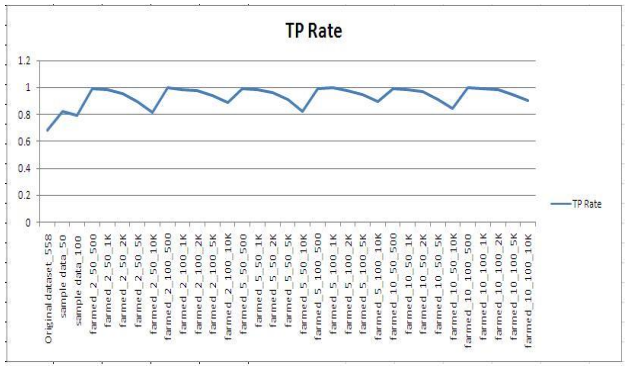

Figure 1. Plot of TP Rate number on instances farmed by the proposed algorithm.

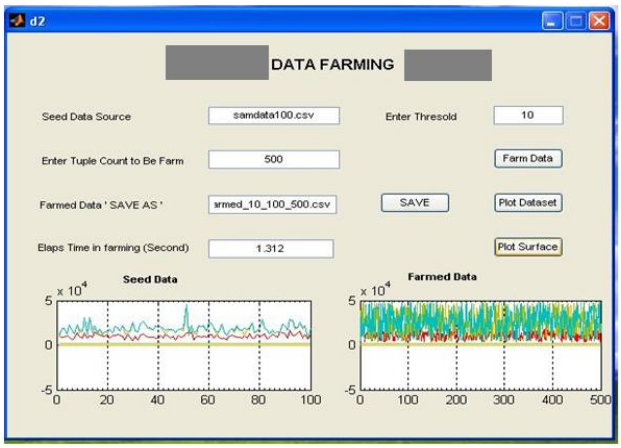

Figure 2. Running Proposed Data Farming Algorithm

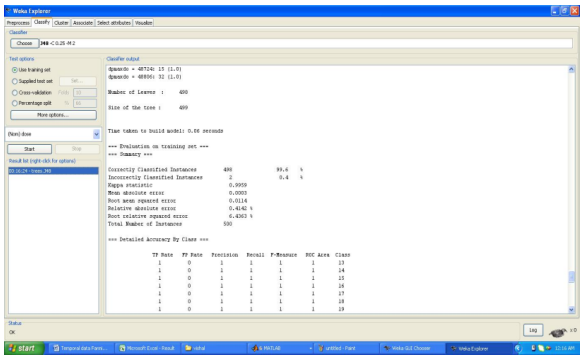

Figure 3. J48 Classification screen shot

Another performance factor is correctly classified Instances (CCI) and Incorrectly Classified Instances (ICI). Correctly classified instances for the original dataset, sample data_50 & sample data_100 are 68.1%, 82% and 79%. And incorrectly classified instances for the original dataset, sample data_50 & sample data_100 are 31.90%, 18% & 21% respectively (see Table-IV). Hence, CCI has increased for the farmed datasets and ICI has decreased. It indicates the farmed data is more appropriate compare to the sample datasets for mining purposes.

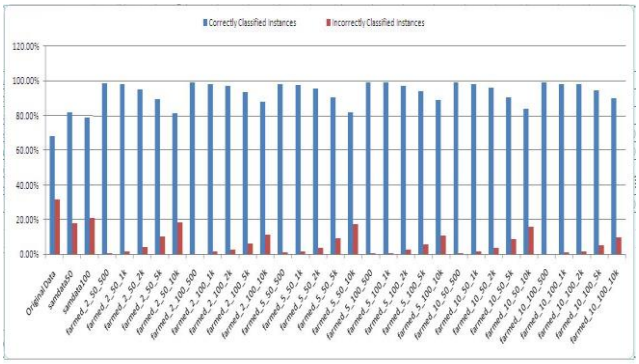

Figure 4. Plot of correctly & incorrectly classified instances by J48 Classification on original, sample & farmed Data

Figure 4 shows the percentage of correctly & incorrectly classified instances for the original, sample, farmed datasets, it can be seen that percentage of correctly classified instances has increased & percentage of incorrectly classified instances has decreased for all the 30 farmed datasets.

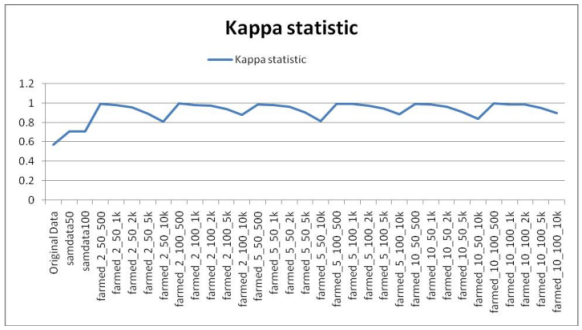

Figure 5. Plot of values of kappa statistics by J48 Classification on original, sample & farmed Data.

TABLE III
J48 CLASSIFICATION RESULTS ON ORIGINAL, SAMPLE & FARMED DATA.

Data Set	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Time
Original dataset_558	0.681	0.032	0.77	0.681	0.7	0.907	0.11
sample data_50	0.82	0.097	0.826	0.82	0.814	0.952	0
sample data_100	0.79	0.036	0.852	0.79	0.798	0.938	0
farmed_2_50_500	0.992	0	0.993	0.992	0.992	1	0.02
farmed_2_50_1K	0.983	0.001	0.984	0.983	0.983	1	0.03
farmed_2_50_2K	0.957	0.002	0.959	0.957	0.957	1	0.22
farmed_2_50_5K	0.896	0.004	0.903	0.896	0.896	0.999	0.8
farmed_2_50_10K	0.816	0.007	0.835	0.816	0.814	0.998	1.88
farmed_2_100_500	0.998	0	0.998	0.998	0.998	1	0.06
farmed_2_100_1K	0.983	0.001	0.984	0.983	0.983	1	0.06
farmed_2_100_2K	0.975	0.001	0.976	0.975	0.975	1	0.25
farmed_2_100_5K	0.938	0.002	0.941	0.938	0.938	1	0.69
farmed_2_100_10K	0.885	0.004	0.893	0.885	0.885	0.999	2.8
farmed_5_50_500	0.988	0	0.989	0.988	0.988	1	0
farmed_5_50_1K	0.982	0.001	0.983	0.982	0.982	1	0.02
farmed_5_50_2K	0.963	0.001	0.964	0.963	0.963	1	0.06
farmed_5_50_5K	0.909	0.003	0.914	0.909	0.909	0.999	0.27
farmed_5_50_10K	0.824	0.007	0.842	0.824	0.822	0.998	2.53
farmed_5_100_500	0.994	0	0.994	0.994	0.994	1	0
farmed_5_100_1K	0.995	0	0.995	0.995	0.995	1	0.02
farmed_5_100_2K	0.975	0.001	0.975	0.975	0.974	1	0.03
farmed_5_100_5K	0.944	0.002	0.946	0.944	0.944	1	0.24
farmed_5_100_10K	0.892	0.004	0.9	0.892	0.891	0.999	1.56
farmed_10_50_500	0.994	0	0.994	0.994	0.994	1	0
farmed_10_50_1K	0.985	0.001	0.986	0.985	0.985	1	0.02
farmed_10_50_2K	0.965	0.001	0.966	0.965	0.964	1	0.05
farmed_10_50_5K	0.909	0.003	0.916	0.909	0.909	0.999	0.25
farmed_10_50_10K	0.841	0.005	0.856	0.841	0.84	0.998	2.56
farmed_10_100_500	0.996	0	0.996	0.996	0.996	1	0.02
farmed_10_100_1K	0.987	0	0.987	0.987	0.987	1	0.02
farmed_10_100_2K	0.984	0.001	0.984	0.984	0.984	1	0.02
farmed_10_100_5K	0.949	0.002	0.951	0.949	0.949	1	0.22
farmed_10_100_10K	0.903	0.003	0.91	0.903	0.903	0.999	0.88

TABLE IV
J48 CLASSIFICATION RESULTS ON ORIGINAL DATASET & SAMPLE DATA OF SIZE 50 & 100.

Name	Factor	Original Data	samdata50	samdata100
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

Kappa statistics is also a measure for the classification accuracy; it has also increased in farmed datasets compare to the original & sample datasets (see figure 5).

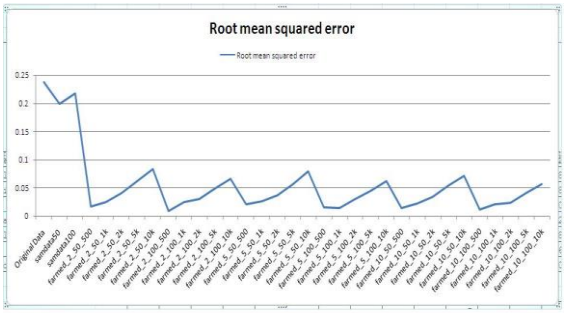

Figure 6. Plot of Root mean squared error by J48 Classification on original, sample & farmed Data.

Root Mean squared error (RMSE) has decreased for the farmed datasets in compare to the original & sample datasets (see figure 6).

Figure 7. Plot of classification result on original, sample & farmed Data.

Figure 7 shows that correctly classified instances (CCI) & kappa statistics (KS) have increased & incorrectly classified instances (ICI), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative absolute error (RAE), Root relative squared error (RRSE) have decreased for the farmed data compare to the original dataset and sample datasets. The time complexity of the proposed algorithm is O (m*n), where m is the number of data to be farmed and n is the number of attribute in the seed dataset. It is quadratic time complexity algorithm.

V. Conclusion

Proposed algorithm farmed the sufficient data with improved adequateness of the available seed dataset for mining. By filling up of missing data & updating predicted values of few attribute we get fertile seed. Proposed algorithm farms more datasets from this fertile seed. 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 compare the sample datasets. Farming time required is highly dependent on the instances to be farm and lightly on the number of seed data & error threshold. correctly classified instances (CCI) & kappa statistics (KS) have increased & incorrectly classified instances (ICI), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative absolute error (RAE), Root relative squared error (RRSE) have decreased for the farmed data compare to the original dataset and sample datasets. This variation shows that the farmed data is more effective compare to the sample datasets.

References

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

[2]. 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. Fowler eds.

[3]. 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.

[4]. Andrew Kusiak, Data Farming Methods for Temporal Data Mining, Intelligent Systems Laboratory, 2139 Seamans Center, The University of Iowa, Iowa City, Iowa 52242 - 1527

[5]. Adam J. Forsyth, Gary E. Horne, Stephen C. Upton, MARINE CORPS APPLICATONS OF DATA FARMING, Proceedings of the 2005 Winter Simulation Conference, M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

[6]. 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.

[7]. 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. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

[8]. http://www.stat.ucla.edu/projects/datasets/cardiac.dat

[9]. Kusiak, A. (2000), Decomposition in data mining: an industrial case study, IEEE Transactions on Electronics Packaging Manufacturing, Vol. 23, No. 4, pp. 345-353.

[10]. Kusiak, A. (2001), Feature transformation methods in data mining, IEEE Transactions on Electronics Packaging Manufacturing, Vol. 24 (in print).

[11]. Horne, G. And Meyer, T. 2004. Data Farming, Briefing Presented at the Informs National Meeting. Denver, Co.

[12]. Simulation Experiments and Efficient Design (SEED) centre for data farming (2009) http:// harvest.nps.edu.

[13]. Gary Horn, Stephen Seichter, Karsten Haymann, Data Farming in Support of Military Decision Makers 2010. http:// harvest.nps.edu.

[14]. Gary Horne, Ted Meyer Data Farming and Defense Applications, Naval Postgraduate School, gehorne@nps.edu, temeyer@nps.edu

[15]. Mohd Shahnawaz, Analysis of Data Farming Methods, 3rd National Conference on Emerging Trends in Software Engineering & Information Technology ETSEIT-2009 on 21-22 March 2009 GEC, Gwalior.

[16]. Mohd Shahnawaz & Kanak Saxena, Analysis of Missing Value Estimation Algorithms for data Farming. International Journal of Engineering Sciences Special issue Sep-2011, Vol. 4, ISSN: 2229-6913.pp 496- 504.

[17]. Mohd Shahnawaz & Kanak Saxena, A Comparative Study of Various Regression Model for Data Farming, International Journal of Wisdom Based Computing (IJWBC) Vol 2(1), 2012 ISSN 2231-4857.pp 29-34.

[18]. Mohd Shahnawaz & Kanak Saxena, A Temporal Data Farming using Iterative Prediction on HDD, International Journal Of Scientific & Engineering Research, Volume 3, Issue 4, May-2012 ISSN 2229-5518.

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

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

[21]. Mohd. Shahnawaz & Kanak Saxena, Data Farming Algorithm with Temporal Medical Events, International Journal of Data Mining and Emerging Technologies, Vol.-5 Issue -1 May 2015 pp 31-37.