



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 Issue: VI Month of publication: June 2020

DOI: <http://doi.org/10.22214/ijraset.2020.6214>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

Increasing the Accuracy of Data Farming Algorithm using Seed Data Set

Ashutosh Anand¹, Dinesh Kumar Sahu², Varsha Namdeo³

^{1, 2, 3}SRK University, Bhopal M.P., India

Abstract: Data farming is a manner of growing enough facts with the help of diverse statistical and heuristic techniques. As statistics collection value is high, such a lot of times records mining projects uses existing records accrued for various other purposes, such as every day accrued facts to technique and records required for monitoring & control. Sometimes, the dataset to be had might be big or wide dataset and sufficient for extraction of understanding but every now and then the dataset is probably slender and insufficient to extract significant understanding or the facts may not even exist. We can cultivate the data where we've restricted informational index and afterward apply the statistics mining calculation to remove the helpful data. We proposed an set of rules for statistics farming steps statistics plantation & harvesting. We farm enough records from the to be had little seed records by making use of the proposed set of rules of information farming. Classification results of J48 class, for farmed statistics is achieved higher than type results for the seed facts, which proves that the proposed facts farming set of rules has produced effective facts. In this thesis, we gift an set of rules for records farming which farms the records with the assist of the seed facts on a predefined errors threshold rate. Proposed set of rules has implemented on farmed datasets are tested for the category accuracy at the weka open source statistics mining tool.

Keywords: J48, WEKA, matlab, data mining, data farming.

I. INTRODUCTION

Information mining is the procedure to extricate significant concealed data from huge database. It is another and developing field of research. It is valuable for the top level chiefs to dissect the past and anticipate what's to come. Troughs can take right choice with the assistance of information mining examination. A short presentation of information mining or information disclosure from information is given here. This field is a blend of different fields like database, measurements, man-made brainpower and activity inquire about and so forth. Essential objective of information mining is to concentrate of revealing intriguing concealed information designs in enormous informational collections.

Information mining is a loosely characterized field; its definition is changed creator to creator. It's for the most part relies upon the foundation and perspectives on the definer. There are a definitions taken from the accessible writing as underneath. It very well may be resolved that information mining is the need of the present market situation. Plato said that-Necessity is the mother of creation. Information mining benefits the general public in different application territories like dynamic, social insurance, safeguard and promoting and so on.. This expansion in the information volume is additionally one of the noteworthy purposes behind the prominence of information mining and information disclosure. The data and information removed or picked up can be utilized for applications like market investigation, clinical human services, and dynamic. Some other application likewise utilized mined information like science investigation, extortion discovery, and client maintenance and creation control.

A. Data Farming

Sufficient information is required for choices making based on information extricated by the information mining process, information assortment is a vital procedure, ordinarily information isn't satisfactory for the mining. All things considered information cleaning, information decrease, choice and information cultivating procedures are applied to get satisfactory information. Subsequent to getting the sufficient information, somebody can apply the mining calculations to remove increasingly exact and helpful data contrasted with the previous information. Techniques and instruments are required for deciding the most suitable information at an adequate expense. As indicated by the past experience, we can see that the information cultivating exertion regularly exceeds the information mining task, particularly in the business. One of the significant explanations for this is the modern information is gathered for reasons other than dynamic. At some point gathered information is a wide scope of highlights that go past conventional models.

Because of the lacking of examination devices, limited or deficient attention to information mining and information cultivating instruments expands the information assortment cost. Information cultivating improves the information available and furthermore decides the most pertinent information to be gathered.

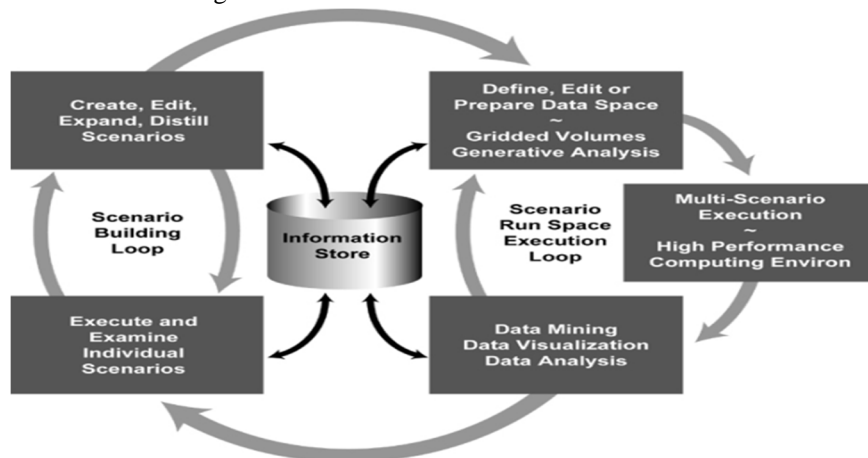


Figure1: Data Farming Approach for High Performance Computing

B. Problem Formulation

Information is developing in the exponential request in last 3-multi decade; thus, the information extraction and examination is getting troublesome. Anyway there exists different information mining calculation and programmed instruments for this reason, one furthest edge of this situation is that, where information isn't accessible in the adequate volume we can't remove the valuable information. For this situation, we can't accomplish the quality data or information from the information. To determine such circumstance a productive calculation for information cultivating is required. Information cultivating is the procedure to become the datasets, like developing yields in horticulture. Information cultivating steps are information preparation, information development, information ranch and information collecting.

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.

- 1) Maximize performance measure (e.g., classification accuracy, cluster density, rule support and confidence)
- 2) 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.

C. Proposed Algorithm

To perform the analysis on the various regression models for data farming steps like data fertilization & data cultivation, we used weka tool. In this section a brief introduction of this tool is given. It's a tool which implements near about all data mining & statistical test algorithms in java programming language.

- 1) *Data Cultivation* – Getting lower bound and upper bound of the scope of every property of seed dataset and applying limit.
- 2) *Data Plantation* – Apply the information developing component as calculation given in pseudo code in this part.
- 3) *Data Harvesting* – Collecting cultivated information and putting away it as the perpetual information storehouse.

D. Process Of Method

Data farming is used to improve the results in terms of these performance measures

- 1) identify and select the key properties, which to be anticipated/assessed/registered.
- 2) Apply k-implies grouping technique on the seed information and discover the bunch.
- 3) Apply relapse on each group datasets and discover the relapse condition (let 3 bunches are produced), at that point we get 3 relapse conditions, where K_{IJ} is the coefficient of X_J for I th group.

$$Y_1 = K_{11}X_{11} + K_{12}X_{12} + K_{13}X_{13} + \dots + K_{1N}X_{1N}$$

$$Y_2 = K_{21}X_{21} + K_{22}X_{22} + K_{23}X_{23} + \dots + K_{2N}X_{2N}$$

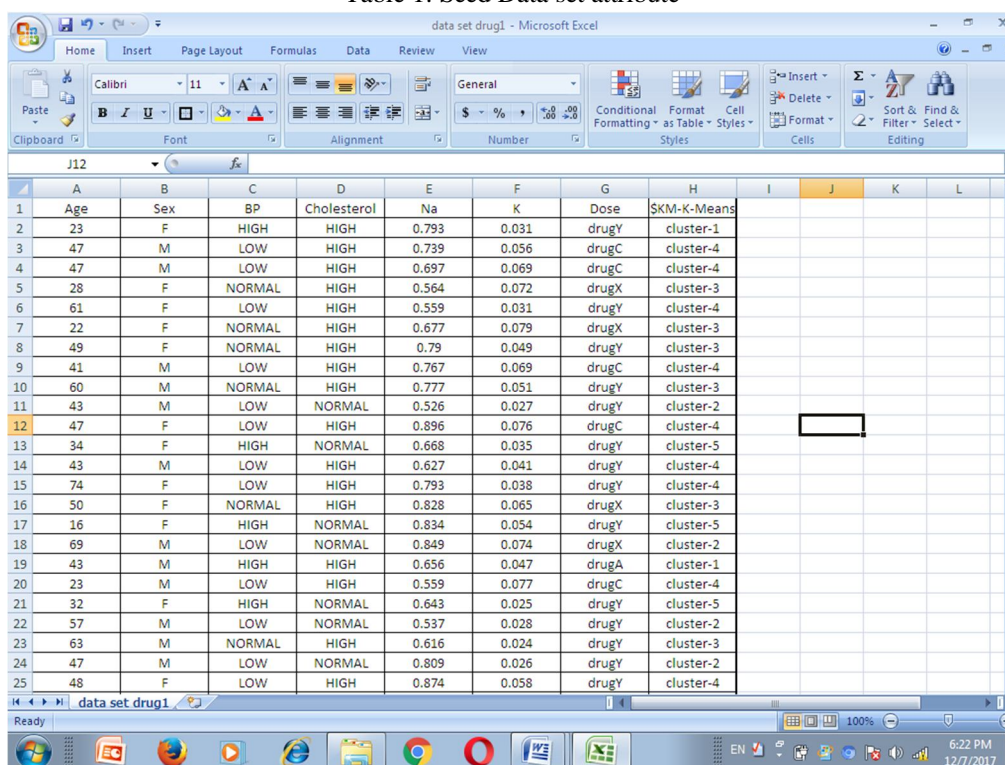
$$Y_3 = K_{31}X_{31} + K_{32}X_{32} + K_{33}X_{33} + \dots + K_{3N}X_{3N}$$

- 4) Apply Prediction on the key characteristic and discover the anticipated estimations of the key quality.
- 5) Now we get the informational index as $(X1, X2, X3... XN, Y, Y', Y'')$. where Y, Y', Y'' are the first second and third occasions anticipated estimations of key quality. Apply again relapse on the dataset $(X1, X2, X3... XN, Y, Y', Y'')$ and discover again relapse condition for each group dataset we get relapse condition as –
 - a) $Y'1 = M11X11 + M12X12 + M13X13 + ... M1NX1N$
 - b) $Y'2 = M21X21 + M22X22 + M23X23 + ... M2NX2N$
 - c) $Y'3 = M31X31 + M32X32 + M33X33 + ... M3NX3N$
- 6) Generate Mutated dataset based on condition 1 and 2 (Increase or lessening XI as the expansion or diminishing in their co-proficient KIJ and MIJ).
- 7) Find out the scope of XI; Hence discover the min and max estimations of XI for each group (for example in the dataset for group one. Estimations of X1 change from 3 to 7, Values of X2 shift from 5 to 10 and Values of X3 differs from 1 to 5).
- 8) Prepare the principles for the group with the assistance of min-max esteems as
- 9) On the off chance that $(3 < X1 < 7)$ AND $(5 < X2 < 10)$ AND $(1 < X3 < 5)$ at that point group 1
- 10) On the off chance that $(1 < X1 < 3)$ AND $(11 < X2 < 15)$ AND $(5 < X3 < 7)$ at that point group 2
- 11) On the off chance that $(7 < X1 < 9)$ AND $(1 < X2 < 5)$ AND $(7 < X3 < 8)$ at that point group 3

II. EXPERIMENTAL RESULT ANALYSIS

To perform the analysis on the various regression models for data farming steps like data fertilization & data cultivation, we used weka tool. In this section a brief introduction of this tool is given. It's a tool which implements near about all data mining & statistical test algorithms in java programming

Table 1: Seed Data set attribute



	A	B	C	D	E	F	G	H
1	Age	Sex	BP	Cholesterol	Na	K	Dose	SKM-K-Means
2	23	F	HIGH	HIGH	0.793	0.031	drugY	cluster-1
3	47	M	LOW	HIGH	0.739	0.056	drugC	cluster-4
4	47	M	LOW	HIGH	0.697	0.069	drugC	cluster-4
5	28	F	NORMAL	HIGH	0.564	0.072	drugX	cluster-3
6	61	F	LOW	HIGH	0.559	0.031	drugY	cluster-4
7	22	F	NORMAL	HIGH	0.677	0.079	drugX	cluster-3
8	49	F	NORMAL	HIGH	0.79	0.049	drugY	cluster-3
9	41	M	LOW	HIGH	0.767	0.069	drugC	cluster-4
10	60	M	NORMAL	HIGH	0.777	0.051	drugY	cluster-3
11	43	M	LOW	NORMAL	0.526	0.027	drugY	cluster-2
12	47	F	LOW	HIGH	0.896	0.076	drugC	cluster-4
13	34	F	HIGH	NORMAL	0.668	0.035	drugY	cluster-5
14	43	M	LOW	HIGH	0.627	0.041	drugY	cluster-4
15	74	F	LOW	HIGH	0.793	0.038	drugY	cluster-4
16	50	F	NORMAL	HIGH	0.828	0.065	drugX	cluster-3
17	16	F	HIGH	NORMAL	0.834	0.054	drugY	cluster-5
18	69	M	LOW	NORMAL	0.849	0.074	drugX	cluster-2
19	43	M	HIGH	HIGH	0.656	0.047	drugA	cluster-1
20	23	M	LOW	HIGH	0.559	0.077	drugC	cluster-4
21	32	F	HIGH	NORMAL	0.643	0.025	drugY	cluster-5
22	57	M	LOW	NORMAL	0.537	0.028	drugY	cluster-2
23	63	M	NORMAL	HIGH	0.616	0.024	drugY	cluster-3
24	47	M	LOW	NORMAL	0.809	0.026	drugY	cluster-2
25	48	F	LOW	HIGH	0.874	0.058	drugY	cluster-4

Algorithm generates data according to the range of the input seed data. Proposed data farming methodology completes in these steps:

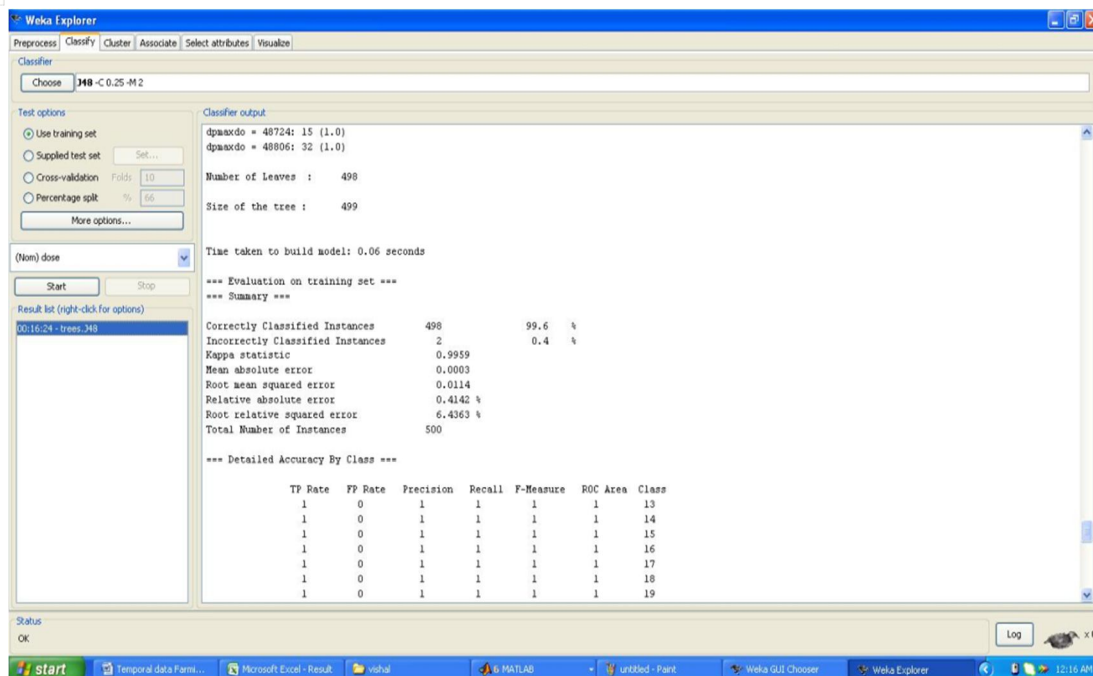


Figure 2: Plot of farmed Data by the propos Algorithm

Figure 2 depicts the running snapshot of the J48 Classification on weka software. we can see the result of the J48 Classification algorithm produced by the weka. There are various parameters obtained by the experiments like correctly classified instances, incorrectly classified instances, kappa statistics, mean absolute error, root mean squared error, relative absolute error & root relative squared error.

Table 2: J48 Classification Result on farmed data on error threshold 5 & seed tuple 50

Name	farmed_5_50_500	farmed_5_50_1k	farmed_5_50_2k	farmed_5_50_5k	farmed_5_50_10k
CCI	98.80%	98.20%	96.25%	90.90%	82.36%
ICI	1.20%	1.80%	3.75%	9.10%	17.64%
KS	0.9875	0.9813	0.9611	0.9056	0.817
MAE	0.0009	0.0013	0.0027	0.0065	0.0127
RMSE	0.0207	0.0254	0.0366	0.0571	0.0797
RAE	1.25%	1.87%	3.89%	9.45%	18.43%
RRSE	11.16%	13.67%	19.72%	30.74%	42.93%
INSTANCE	500	1000	2000	5000	10000

Table 2 contains the results obtained from the weka software by applying J48 classification on permissible threshold value 5, seed data size 50 and farmed tuple 500, 1 k, 2 k, 5 k and 10 k. Incorrectly classified instances are 1.20, 1.80, 3.75, 9.10 and 17.64% respectively while kappa statistics are 0.9875, 0.9813, 0.9611, 0.9056 & 0.817 respectively.

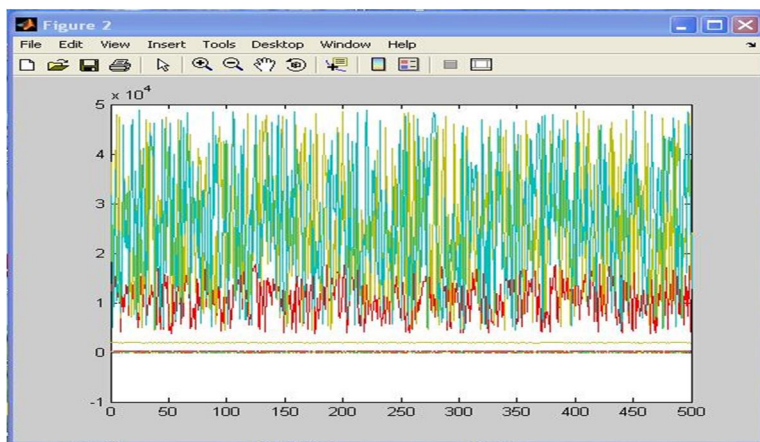


Figure 3: Plot of farmed Data by the proposed Algorithm

Figure 3: shows the farmed data in graphical view. In this figure 500 tuples are farmed from the seed data set by the proposed algorithm.

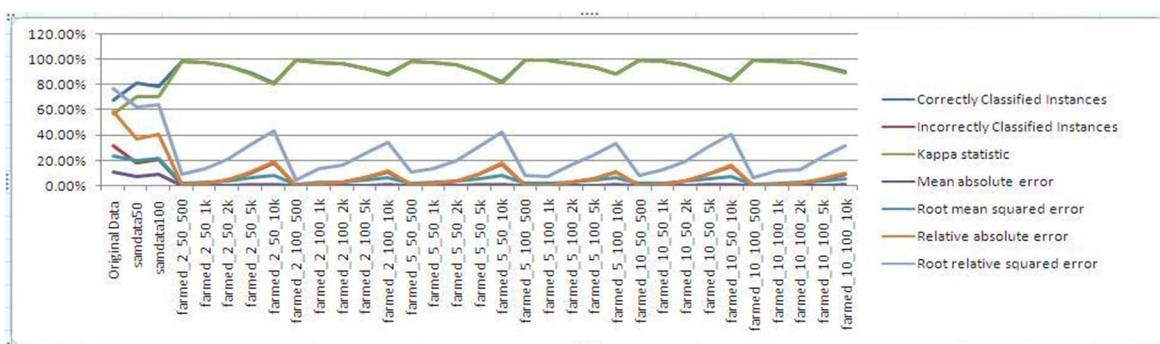


Figure 4: Plot of classification result on original, sample & farmed Data

Figure 4 shows that accurately characterized examples (CCI) and Kappa measurements (KS) are expanded and inaccurately arranged occurrences (ICI), Mean supreme mistake (MAE), Root mean squared blunder (RMSE), Relative outright blunder (RAE), Root relative squared blunder (RRSE) are diminished for the cultivated information contrasted with the first dataset and test datasets. The time multifaceted nature of the proposed calculation is $O(mn)$, where m is the quantity of information to be cultivated and n is the quantity of traits in the seed dataset. It is a quadratic time multifaceted nature calculation.

III. CONCLUSION AND FUTURE WORK

Data farming is an emerging discipline of research inside the present day scenario, where facts collection fee and time ate up in data collection are vast to reduce. We proposed an algorithm for statistics farming steps statistics plantation & harvesting. We farm sufficient information from the available little seed facts by means of applying the proposed algorithm of information farming. Proposed set of rules farmed sufficient information with progressed adequateness of the to be had seed dataset for mining. By filling up of missing data & updating expected values of few attributes, we get fertile seed dataset & via cultivation we prepare the environment for plantation. Proposed set of rules plant life those fertile seed in a cultivated environment & harvests the crops in the form of farmed facts. We can see that the farmed records is sufficient to perform diverse mining strategies and find out the hidden knowledge whilst seed information is not sufficient. Classification accuracy of the farmed statistics proved, that it is better as compared to the sample datasets. Farming time required is exceedingly depending on the instances to be farmed and lightly on the quantity of seed statistics & mistakes threshold. Correctly classified instances (CCI) & kappa statistics (KS) are increased & incorrectly classified times (ICI), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative absolute mistakes (RAE), Root relative squared errors (RRSE) are reduced for the farmed statistics when as compared to the unique dataset and pattern datasets. This variation suggests that, the farmed facts is more powerful compared to the pattern datasets.

This thesis provides the overall end of the studies work finished on this thesis as well as limitations and future scope of the work. This can be helpful in the further studies in this subject. The present day work may be enhanced inside the destiny with the idea of cloud computing surroundings.

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 2018 Winter Simulation Conference S. J. Mason, R. R. Hill, L. Monch, O. Rose, T. Jefferson, J.W. Fowlereds. 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, Schwier, Data Farming Around the World Overview, Proceedings of the 2015 Winter Simulation Conference, S. J. Mason, R. R. Hill, L. Monch, O. Rose, T. Jefferson, J.W. Fowlereds. 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, eds.
- [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.pdf>
- [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 2017, SVR-Based Data Farming Technique for Web Application. In Ifip International Federation 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] Han J, Kamber M 2001 Data Mining: Concepts and Techniques (San Francisco, CA: Morgan Kaufmann) <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 2015 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 2018 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.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)