International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: http://doi.org/10.31695/IJASRE.2018.32936

Volume 4, Issue 11
November - 2018

Study of Assigning Imperfect Attribute Values for Classifier

Thi Thi Soe[1] and Zarni Sann[2]

[1]Faculty of Computer Science,

[2]Faculty of Computer Systems and Technologies

University of Computer Studies (Mandalay)

Myanmar

ABSTRACT

One of the interesting and important fields of research in data mining is classification on imperfect data. Unknown value can appear in real-world data sets at the stage of data collection. These facts lead to the imperfection of the decision system. The attention of this paper is to study the methodologies in making a decision point on such incompletion of data sets. We found a well-known rough set (RS) based classification scheme: Learning from Examples based on Rough Sets (LERS) that could be treated missing data, numeric data, and inconsistent data set. In this study, we utilize two interpreted meaning of imperfect values: lost values and attribute-concept values. The classification system is illustrated using a case study of iris dataset from the UCI repository. The system is intended to present a comprehensive view of assigning on imperfect attributes value that generates better result among lost and attribute-concept values.

Keywords: *Data mining, Imperfect data, lost value, Attribute–concept value, Rough set-based Classifier.*

1. INTRODUCTION

Mathematical rough set theory was first appeared in the pioneering work in [1]. The idea of rough set was successfully implemented and used in many applications such as classification, pattern recognition, learning algorithms. Classical rough set can only deal with complete information systems in which the data available. However unknown values can attach in real world data at the stage of data gathering. Imprecise and imperfection of data table may affect the performance of pattern analysis and classification system. Hence concentration of many researchers turned to develop algorithms for attenuate such type of problems. Large bodies of attempts extending the rough set have been demonstrated in knowledge acquisition by discovering rules set when presented a datasets with missing data. Among them, the most successful one is system LERS introduced in [2]. LERS applied the algorithms: Learning from Examples Module version1 and version2, LEM1 and LEM2 [2, 3, 4]. Global and local coverings were used in LEM1 and LEM2 respectively. For decision rule generation, LEM2 algorithm was successfully implemented and used in [5-6]. Other approaches using LEM2 algorithm were described in [7-10].

Because of a major inconvenience in LEM2 on processing numerical attribute values, an approach is extended on it [11]. This extended version is called modifications of LEM2 or simply MLEM2. Later several researches were further presented to the fields of analyzing numerical data sets and missing attribute values [12-14]. LERS could also be manipulated inconsistent data set in which two or more cases with all of same attribute values have different decision value. Until current time the researcher continuously analyzing the data mining fields by utilizing system LERS. For example, in resent research [15] comparison results of mining incomplete and inconsistent data were presented in terms of an error rate. Experiments were conducted on 204 data sets and used rule induction of MLEM2 algorithm. They concluded that incompleteness was worse than inconsistency for data mining. Missing type lose values were better than “do not care conditions”. Hence RS-based classifier is leading to a fruitful research and applications in many places. For comparison purpose an RSFit approach was introduced in [16] to assign missing attribute values from rough sets perspective. They demonstrated the RSFit approach by an artificial car data set in [17], UCI data sets and a Geriatric care data set with randomly selected missing attribute values. The accuracy of the prediction was compared to the “closest fit” approach proposed by Grzymala-Busse. They point out that the RSFit approach significantly reduced the computation time and achieved comparable accuracy result.

Generally, the methods for incomplete data classification based on attributes require extra computing: imputation or classifier updating. Imputation is a procedure that consists of assigning the missing values with predicted ones based on available information in the data set. These methods lead to be computational burden. Reference[18] distinguished the missing data into three types: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR).There have been proposed other approaches on how to manipulate missing data in the literature. Reference [19] proposed an approach, using the entropy to estimate the missing values. Experimental results showed that this approach could be able to outperform other imputation techniques, as mean and mode. In [20] a partial imputation technique was developed. It involves the imputation of missing data using complete objects in a small neighborhood of the incomplete ones. In [21] analyzed the efficiency for missing data treatment using classification-based mining algorithm C4.5 and cluster-based mining algorithm K-means. In these experiments, missing values were artificially imputed, in different rates and attributes, into the numerical data sets. K-means imputation method provided good results in most cases of their experiments.

Support vector machines (SVMs) [22] are learning models frequently used for classification. In [23], the imputation technique using SVR was studied and compared it with the traditional imputation techniques mean, median, the mean of the two closest neighbor values and the value of the nearest neighbor techniques. Their results showed that the SVR technique obtained the highest precision with regards to the others methods. Reference [24] examined the impact of performing missing data imputation with five imputation methods: mean, the Hot deck method, Naïve Bayes, multiple imputation and framework method using Naïve Bayes and Hot deck. Experimental analysis was tested with six modern classifiers (RIPPER, C4.5, KNN, SVM, and Naïve Bayes) on 15 data sets to investigate the effect of imputation on the classification errors. Performed analysis result shows that imputation methods are beneficial except mean imputation for the classification of objects with missing attributes.

According to a brief overview of incomplete data handling techniques discussed in [25], the missing data imputation approaches based on machine learning, artificial neural network algorithms, K-nearest neighbor algorithm and Self-Organizing Maps (SOM) are more frequently used. On the other hand three techniques: Hot deck imputation, mean substitution, and Regression substitution are rarely used due to poor classification performance. Among previous classification methodologies when analyzing datasets with missing data, our preferences are on the series of Grzymala-Busse's data mining tasks [11, 12, 14, 26-29]. Hence the goal pursued by this paper is to create a Java code program for LERS- classification framework. The goal of this paper is also to give a comprehensive view related to the meaning of imperfect attribute values: lost and attribute-concept values by providing implementation results.

2. METHOD AND MATERIAL

2.1 Incomplete Information systems

Dealing with the rough set concept, an information system, I, is defined as a pair $I = (U, A)$, where U is a non-empty finite set of objects (cases) called the universe and A is a non-empty finite set of attributes such that $a : U \to V_a$ for every $a \in A$. The set V_a is called the value set of a . A decision system is special form of information system such that $I = (U, A \cup \{d\})$, where d is the decision attribute. The elements of A are called conditional attributes or simply conditions [30]. Decision tables can be utilized as the training data sets for data mining tasks. It is possible in an information system some positions of attribute values are empty or absent. This called an incomplete information system [31, 32]. Also refers to imperfect decision table as described in Table 1.

Table 1: Imperfect decision table

Case	Attributes				Decision
	Att.1	Att.2	Att.3	Att.4	D
1	a11		a13	a14	d1
2	a21		a23	a24	d2
3	a31	a32	a33		d2

2.2 Meaning of Missing values

Absent positions can be assigned with a value based on the available information of data set. Grzymala-Busse and other researchers [33, 26, 27, 28, 29] have been devoted the meaning of missing attribute values in incomplete decision table. As

DOI: 10.31695/IJASRE.2018.32936

described in Table 2 they defined the definitions of missing types along with the relevant symbols and distinguished how to assign a value on each symbol for successive RS-based data mining tasks.

Table 2: Three types of missing attribute values

Definition	Symbol	Assign type	Assume
lost value	?	not assigned by any existent value from the attribute domain	erase
“Do not care” condition	*	assigned by any possible value from the attribute domain	unnecessary
Attribute-concept value	–	assigned by any possible value from the attribute domain bounded to the decision value	has a concept

2.3 Classifier−LERS

Rule induction is one of the important steps in machine learning classification. The basic rule induction algorithm for classifier-LERS is MLEM2. MLEM2 could be able to discover rules from numerical and symbolic attribute of imperfect decision table [11].

2.3.1 Transforming numeric-values

To induce rules from decision table, the first step is discretization of numerical attributes. Doing so required the transformation processes listed below [12].

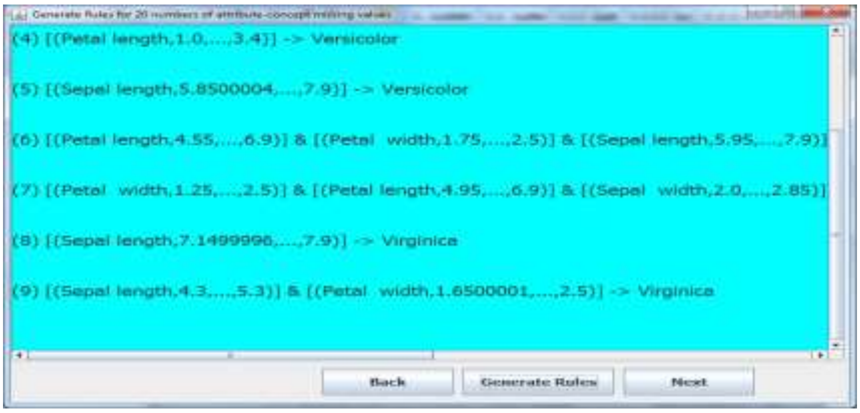
- sorts all numeric values excluding missing attribute values
- computes cut-points as averages for any two consecutive values of the sorted list
- creates two blocks : all cases with numeric value $<C$, and all cases with numeric value $>C$

2.3.2 Central blocks

Central blocks: the attribute-value pair block is the direct basic for RS-based classifier. The attribute-value pair block $[(a, v)]$ is defined as the set $\{x \in U| \ a(x) = v\}$ [34]. For imperfect decision tables, the definition of the block is modified according to the missing types described in Table 2. The block is used to derive the characteristics relation, characteristics set, and lower and upper approximations. We found that imperfect data sets are usually analyzed using three classes of approximations such as singleton, subset and concept [12, 6]. In our implementation, we used the concept definition of A -lower and A -upper approximation for discovering certain and possible rules set with equations, $\underline{A}^{concept}(X) = \cup \{K_A(x)|x \in X, K_A(x) \subseteq X\}$ and $\overline{A}^{concept}(X) = \cup \{K_A(x)|x \in X, K_A(x) \cap X \neq \emptyset\} = \cup \{K_A(x)|x \in X\}$, where X is a subset of U, and$K_A(x)$, the characteristic set is the intersection of the sets $K(x, a)$, for all $a \in A$. For classify testing data and unknown cases, classifier-LERS utilizes the set of rules induced by the algorithm [14, 12].

3. IMPLEMENTATION

In order to illustrate the study area we create Java code program on iris plant dataset from UCI database [35]. The data set contains 150 cases in which 50 in each of three classes. The process flow of classification system on imperfect data is followed by LERS- classifier [12, 11, 14]. In this experiment, the data set is divided into 120 cases for training and 30 cases for testing according to the holdout method. Training data sets are built by randomly placing he lost values at 20 and 40 positions for representing the imperfection of data. Attribute-concept values are also placed at the same positions. The blocks derive as a vital engine at the initial state of computing characteristic set until generating the certain and possible rules based on the concept lower and upper approximation. Rules generated by MLEM2 algorithm for 20 numbers of ‘attribute-concept’ missing values are presented to the user as shown in Figure 1.

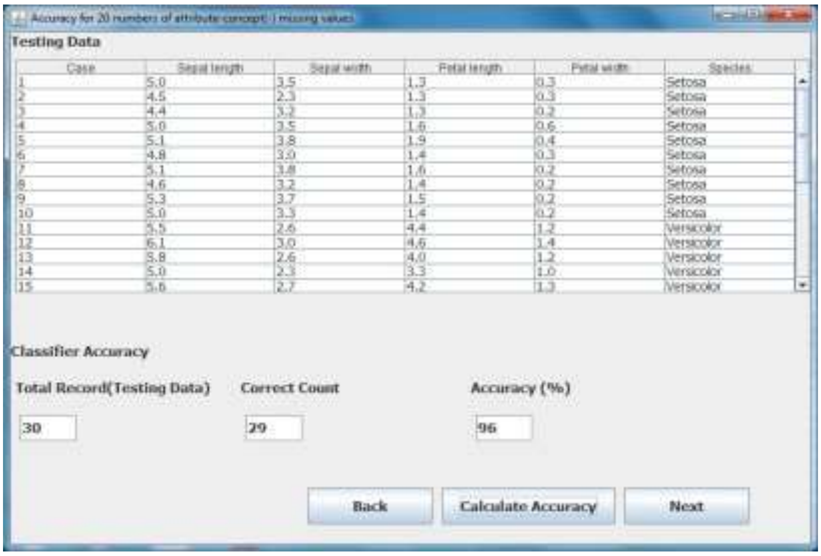

Generate Rules for 20 numbers of attribute-concept missing values
Back
Generate Rules
Next

```
(4) [(Petal length,1.0,...,3.4)] -> Versicolor

(5) [(Sepal length,5.8500004,...,7.9)] -> Versicolor

(6) [(Petal length,4.55,...,6.9)] & [(Petal  width,1.75,...,2.5)] & [(Sepal length,5.95,...,7.9)]

(7) [(Petal  width,1.25,...,2.5)] & [(Petal length,4.95,...,6.9)] & [(Sepal  width,2.0,...,2.85)]

(8) [(Sepal length,7.1499996,...,7.9)] -> Virginica

(9) [(Sepal length,4.3,...,5.3)] & [(Petal  width,1.6500001,...,2.5)] -> Virginica
```

Figure 1: Generated rules for 20 numbers of ‘attribute-concept’ missing values

These rules are employed by LERS classifier to classify thirty cases of testing data. The quality of classifier is measured using the following equation: Testing data and the result of calculated accuracy is given in Figure 2.

$$\text{classification accuracy} = \frac{numberofcorrectlyclassifiedcases}{totalnumberoftestingcases} \times 100 \qquad (1)$$



Accuracy for 20 numbers of attribute-concept(-) missing values

Testing Data

Case	Sepal length	Sepal width	Petal length	Petal width	Species
1	5.0	3.5	1.3	0.3	Setosa
2	4.5	2.3	1.3	0.3	Setosa
3	4.4	3.2	1.3	0.2	Setosa
4	5.0	3.5	1.6	0.6	Setosa
5	5.1	3.8	1.9	0.4	Setosa
6	4.8	3.0	1.4	0.3	Setosa
7	5.1	3.8	1.6	0.2	Setosa
8	4.6	3.2	1.4	0.2	Setosa
9	5.3	3.7	1.5	0.2	Setosa
10	5.0	3.3	1.4	0.2	Setosa
11	5.5	2.6	4.4	1.2	Versicolor
12	6.1	3.0	4.6	1.4	Versicolor
13	5.8	2.6	4.0	1.2	Versicolor
14	5.0	2.3	3.3	1.0	Versicolor
15	5.6	2.7	4.2	1.3	Versicolor

Classifier Accuracy

Total Record(Testing Data): 30

Correct Count: 29

Accuracy (%): 96

Figure2: Classifier accuracy for 20 numbers of 'attribute-concept' missing values

And then, the system tested with 40 numbers of attribute-concept values, 20 and 40 numbers of lost values as in similar ways. The performance of the classifier is reported according to the number of imperfect data for each missing types. Figure 3 shows the classifier accuracy with the bar graph representation for comparison purpose.

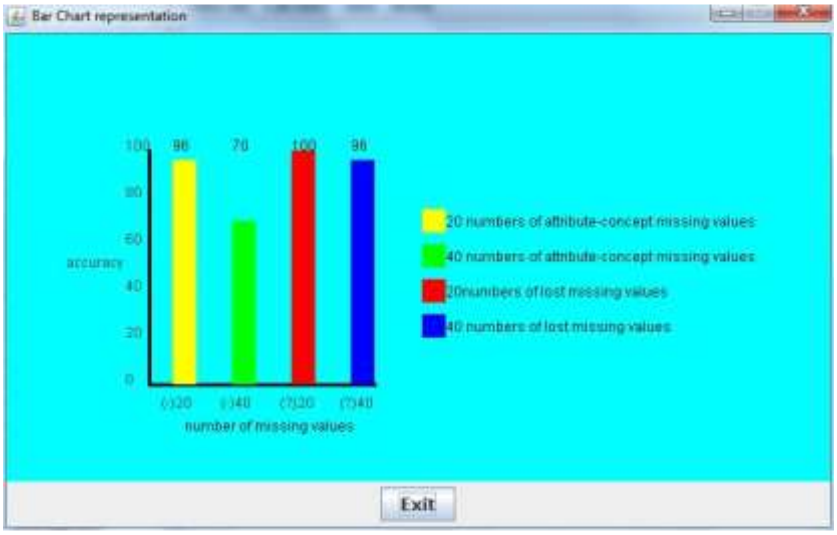


Figure3: Comparisons of classifier accuracy for different missing types

The experiment also reports the generated number of rules and conditions tested on each of the missing types. Results are tabulated in Table 3.

Table 3: Comparison result for lost and attribute-concept values

Number of missing value	Lost		Attribute-concept	
	number of rules	number of conditions	number of rules	number of conditions
20	9	18	9	20
40	10	24	6	11

DOI: 10.31695/IJASRE.2018.32936

4. CONCLUSION

This paper studied the methods of assigning on imperfect and uncertain data for classification task. For treating absent and uncertain data, rough set- based classifier provides as a great mechanism. Concerning with the study area in this paper we implemented the classifier which is based on the operations of LERS- system. This paper also analyzed the classifier accuracy by assigning values on imperfect data deals with the description of lost and attribute-concept values. Classifier utilizing lost values gained the better accuracy than the attribute-concept values both for different numbers of imperfect data: 20 and 40. The resulted accuracy rates also meet the solution of classification on incomplete data by previous works. Obtaining knowledge on studied domain classification model will be constructed combining rough set and other approaches in the future.

REFERENCES

[1.] Z. Pawlak, “Rough sets”, International Journal of Computer and Information Sciences, 1982, Vol. 11, No.5,341–356.

[2.] J. S. Dean, J. W Grzymala-Busse, “An overview of the learning from examples module LEM1”, Technical report, Department of Computer Science, University of Kansas, 1988.

[3.] J. W Grzymala-Busse, “LERS—A System for Learning from Examples based on Rough Sets”, in: Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory (R. Slowinski, Ed.), Kluwer Academic Publishers, Dordrecht, Boston, London, 1992, 3–18.

[4.] C. C. Chan, J. W Grzymala-Busse, “On the attribute redundancy and the learning programs ID3, PRISM, and LEM2”, technical report, Department of Computer Science, University of Kansas, 1991.

[5.] An, A., Cercone, N. “ELEM2: A Learning System for more Accurate Classifications”, Proceedings of the 12th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, 1998.

[6.] J. W Grzymała-Busse, “Rough Set Strategies to Data with Missing Attribute Values”, Proc. of the workshop on Foundations and New Directions in Data Mining. Melbourne, 19-22 November 2003, 56–63.

[7.] J. G Bazan, M. S Szczuka, A Wojna, M. Wojnarski, “On the Evolution of Rough Set Exploration System”, Proceedings of the Rough Sets and Current Trends in Computing Conference, 2004.

[8.] J. W Grzymała-Busse, S.Siddhaye, “Rough Set Approaches to Rule Induction from Incomplete Data”, Proc. of the IPMU2004. Perugia, Italy, 4-9 July 2004, Vol.2, 923–930.

[9.] J. W. Grzymala-Busse, J.Stefanowski, S. Wilk, “A comparison of two approaches to data mining from imbalanced data”, Proceedings of the 8-th International Conference on Knowledge-Based Intelligent Information & Engineering Systems, (KES 2004), 2004.

[10.] J.W. Grzymala-Busse, M. Hu, “A Comparison of Several Approaches to Missing Attribute Values in Data Mining”. W. Ziarko and Y.Yao (Eds.): RSCTC 2000, LNAI 2005 (2001) 378–385.

[11.] J. W.Grzymala-Busse, “MLEM2: A new algorithm for rule induction from imperfect data”, Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge Based Systems, IPMU 2002, July 1–5, Annecy, France, 243–250.

[12.] J. W. Grzymala_Busse, “Data with Missing Attribute Values: Generalization of Indiscernibility Relation and Rule Induction”, Springer_Verlag, Berlin, Heidelberg, 2004.

[13.] J. W. Grzymala-Busse, “A comparison of three strategies to rule induction from data with numerical attributes”, in: Proceedings of the International Workshop on Rough Sets in Knowledge Discovery, associated with the European Joint Conferences on Theory and Practice of Software, Warsaw, Poland, April 5–13, 2003, 132–140.

[14.] J. W. Grzymala-busse, “A local version of the MLEM2 algorithm for rule induction”, fundam. informat. 100 (2010) 1–18.

[15.] P. G. Clark, C. Gao, J. W. Grzymala-Busse, “A comparison of mining incomplete and inconsistent data”, International journal of information, technology and control, Vol. 46,No. 2, 2017, pp. 183-193, DOI 10.5755/j01.itc.46.2.17330.

[16.] J Li, and N. Cercone, “Assigning Missing Attribute Values Based on Rough Sets Theory”, IEEE International Conference on Granular computing, 2006.

[17.] X. L. Hu, T., Han, J.: “A New Rough Sets Model Based on Database Systems”. Fundamenta Informaticae 59 no.2-3 (2004) 135–152.

[18.] R. J. Laird, D. B. Rubin, Statistical Analysis with Missing Data. Second Edition. John Wiley and Sons, New York, 2002.

[19.] T. Delavallade and T.H. Dang, “Using Entropy to Impute Missing Data in a Classification Task” In: IEEE International Conference on Fuzzy Systems, London, 2007, 1–6.

[20.] S. Zhang, “Parimputation,: From imputation and null-imputation to partially imputation”, IEEE Intelligent Informatics Bulletin, 2008, vol. 9 (1), pp. 32-38.

DOI: 10.31695/IJASRE.2018.32936

[21.] B. Mehala , P. Ranjit Jeba Thangaiah, and K. Vivekanandan, “Selecting Scalable Algorithms to Deal With Missing Values”, International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009, pp.80–83.

[22.] N. Cristianini and J. Shawe-Taylor, “Support Vector Machines and other kernel-based learning methods”, Cambridge University Press, UK, 2000.

[23.] F. Honghai, C. Guoshun, Y. Cheng, Y. Bingru and C. Yumei, “A SVM Regression Based Approach to Filling in Missing Values”, LNCS - Knowledge-Based Intelligent Information and Engineering Systems, Springer Berlin - Heidelberg, vol. 3683, 2005, 581–587.

[24.] A. Farhangfara, L. Kurganb and J. Dy, “Impact of imputation of missing values on classification error for discrete data”, Pattern Recognition, vol. 41, 2008, 3692–3705.

[25.] N.C. Vinod and M. Punithavalli, “Classification of Incomplete Data Handling Techniques-An Overview”, International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397, Vol. 3 No. 1 Jan 2011, 340–344.

[26.] J. W. Grzymala-Busse, “Three approaches to missing attribute values– rough set perspective”, in Proceedings of the Workshop on Foundation of Data Mining, in conunction with the Fourth IEEE International Conference on Data Mining, 2004, 55–62.

[27.] J. W. Grzymala-Busse and A. Y. Wang, “Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values”, in Proceedings of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97), 69–72.

[28.] J. W. Grzymala-Busse, “On the unknown attribute values in learning from examples”, In Proceedings of the ISMIS-91, 6th International Symposium on Methodologies for Intelligent Systems, 368–377.

[29.] M. Kryszkiewicz, “Rules in incomplete information systems”, Information Sciences, 1999, 113(3-4):271–292.

[30.] Komorowski, J., Polkowski, L., and Skowron, A., “Rough sets: A Tutorial”, Springer_Verlag, 1999.

[31.] S. Greco, B. Matarazzo, R. Stowinski, “Handling missing values in rough set analysis of multi-attribute and multi-criteria decision problems”, Lecture Notes in Computer Science 1711 (1999) 146–157.

[32.] Z. Meng, Z. Shi, “A fast approach to attribute reduction in incomplete decision systems with tolerance relation-based rough sets”, Information Sciences, 179 (2009) 2774–2793.

[33.] J. Stefanowski and A. Tsoukias, “On the extension of rough sets under incomplete information”, in Proceedings of the RSFDGrC'1999, 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing,73-81.

[34.] J. W. Grzymala-Busse, “A new version of the rule induction system LERS”, Fundamenta Informaticae, 1997, 31, 27–39.

[35.] M. Lichman, UCI Machine Learning Repository http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data.

DOI: 10.31695/IJASRE.2018.32936