ISSN:2321-1156

International Journal of Innovative Research in Technology & Science(IJIRTS)

COMPARISON OF TWO DISCRIMINANT ANALYSIS ALGORITHMS IN PESTICIDES TOXICITY CLASSIFICATION

Jyothy S T, RNSIT, Bangalore, Deepu Kumar T L, Tata Elxsi Limited, Bangalore, Dr. Andhe Pallavi, RNSIT, Bangalore

Abstract

The pesticide toxicity levels are found using two discriminant analysis algorithms - Linear Discriminant Analysis (LDA) and Quadrature Discriminant Analysis (QDA) in this paper. LDA is based upon the concept of searching for a linear combination of variables (predictors) that best separates two classes. The basic idea is, for each class to be identified, calculate a (different) linear function of the attributes. The class function yielding the highest score represents the predicted class. The goal of this approach is to identify the toxicity level of the pesticides by LDA. LDA algorithm tells which toxicity level a pesticide will belongs to. In QDA probability density function is obtained with respect to specific co variance matrix. The algorithms tell whether the pesticide will belong to a level ranging from highly toxic level to non toxic level. By knowing the toxicity level of a pesticide, proper pesticide can be used for infected plant. Some plants are not too much affected by insects; these plants require low level toxicity pesticide. Some plants are too much affected by insects; these plants require high toxicity pesticide. Results of both the algorithms are compared to estimate the suitable algorithm for pesticide toxicity classification.

Introduction

Currently study of the consequences of chemicals on the health of human beings and wildlife is performed by ad-hoc experiments, which are very expensive, years long, and involve animal studies [2]. The huge number of compounds to be studied makes this especially challenging. This research area requires new and efficient computer-based approaches to analyse huge and complex amounts of information and to automatically discover and use new knowledge implicitly contained in the data. The goal of toxicity prediction is to describe the relationship between chemical properties, on the one hand, and biological and toxicological processes, on the other. Knowledge about the causes of toxicity is incomplete. No single property can satisfy the requirement to model the toxic activity, with some interesting successful cases. Thus, a large number of parameters are of potential interest. The problem is how to deal with this high dimensional information

Discriminant Analysis (DA), a multivariate statistical technique is used to build a predictive / descriptive model of group discrimination based on observed predictor variables and to classify each observation into one of the groups [3]. In DA multiple quantitative attributes are used to discriminate single classification variable. Thus, it is different from the cluster analysis since DA requires prior knowledge of the classes, usually in the form of a sample from each class. The objectives of DA are i) to investigate differences between groups ii) to discriminate groups effectively; iii) to identify important discriminating variables; iv) to perform hypothesis testing on the differences between the expected groupings; and v) to classify new observations into pre-existing groups.

Linear discriminant analysis (LDA) is one of the oldest mechanical classification systems [4]. Linear Discriminant Analysis (LDA) is a classification method originally developed in 1936 by R. A. Fisher. It is simple, mathematically robust and often produces models whose accuracy is as good as more complex methods. LDA is based upon the concept of searching for a linear combination of variables (predictors) that best separates two classes (targets). A simple linear correlation between the model scores and predictors can be used to test which predictors contribute significantly to the discriminant function. Correlation varies from -1 to 1, with -1 and 1 meaning the highest contribution but in different directions and 0 means no contribution at all.

The basic idea of LDA is simple: for each class to be identified calculate a (different) linear function of the attributes [5]. The class function yielding the highest scores represents the predicted class. There are many linear classification models, and they differ largely in how the coefficients are established. One nice quality of LDA is that, unlike some of the alternatives, it does not require multiple passes over the data for optimization. Also, it naturally handles problems with more than two classes and it can provide probability estimates for each of the candidate classes. Some analysts attempt to interpret the signs and

magnitudes of the coefficients of the linear scores, but this can be tricky, especially when the number of classes is greater than 2. LDA bears some resemblance to principal components analysis (PCA), in that a number of linear functions are produced (using all raw variables), which are intended, in some sense, to provide data reduction through rearrangement of information. Note, though, some important differences: First, the objective of LDA is to maximize class discrimination, whereas the objective of PCA is to squeeze variance into as few components as possible. Second, LDA produces exactly as many linear functions as there are classes, whereas PCA produces as many linear functions as there are original variables. Last, principal components are always orthogonal to each other ("uncorrelated"), while that is not generally true for LDA's linear scores.

Quadrature Discriminant Analysis is one of the discriminant analysis classification algorithms. In QDA probability density function obtained with respect to specific co variance matrix is used.

Pesticides data containing multi-attributes is used to demonstrate the features of linear discriminant analysis in discriminating the four toxic groups,

- HIGH
- MODERATE
- LOW
- NON TOXIC

Class 1 represents "HIGH TOXIC GROUP", Class 2 represents "MODERATELY TOXIC GROUP", Class 3 represents "LOW TOXIC GROUP", and Class 4 represents "NON TOXIC GROUP".

Data Set

In this paper, a data set constituted of 45 common organophosphorous compounds has been investigated. The toxicity value was expressed using the form Log10 (1/LC50). Then the values were scaled in the interval [-1...1]. Four classes were defined: Class 1 [-1..-0.5), Class 2 [-0.5...0), Class 3 [0...0.5), Class 4 [0.5...1].

45 organophosphorous compounds and true class for compounds are shown in Table 1 [1].

Table 1. Compounds and their true classes

Compound	True Class
Anilofos	2
Chloropyrifos	1
Isazofos	1
Phosalone	2
Prothiofos	2
Azamethiphos	2
Azinphos-methyl	1
Diazinon	3
Phosmet	2
Pirimiphos-ethyl	1
Pirimiphos-methyl	2
Pyrazophos	2
Quinalphos	1
Azinphos-ethyl	1
Etrimfos	1
Fosthiazate	4
Methidathion	1
Piperophos	3
Triazophos	1
Dichlorvos	2
Disulfoton	3
Fenamiphos	4
Fenthion	2
Fonofos	1
Isofenphos	3
Methamidophos	4
Omethoate	3
Parathion	2
Parathion-methyl	3
Phoxim	2
Sulfotep	1
Tribufos	2
Trichlorfon	2
Acephate	4
Dimethoate	3
Ethion	2
Ethoprop	3
Fenitrothion	3
Formothion	3
Phorate	1
Propetamphos	3

Sulprofos	3
Temefos	3
Terbufos	1
Thiometon	3

For each compound eight descriptors are analysed. Descriptors used for analysis and toxicity classification shown in table 2 [6].

Table 2. Descriptors used for analysis and toxicity classification

Descriptors	Unit	Code
Melting Point	deg	D1
Log P	N/A	D2
water Solubility	mg/L	D3
Vapour Pressure	mm Hg	D4
Henry's Law Constant	atm-m3/mole	D5
Atmospheric OH Rate Constant	cm3/molecule-sec	D6
Toxicity for rat(oral)	mg/kg	D7
Toxicity for rat(skin)	mg/kg	D8

Simulation

Pesticide toxicity level is analyzed using LDA and QDA algorithm, which is implemented using MATLAB. Results of the LDA and QDA algorithm are analyzed with graphical representation.

The implementation carried in MATLAB is described in the steps below [7].

1. Compounds belonging to same class are stored in one matrix. Since there are four classes, four matrices are used to store similar compounds properties in x1, x2, x3 and x4 respectively.

2. Mean for each class and whole data set containing all classes are calculated. Mean is calculated by adding all the values for each descriptor and this value is divided by total number of compounds in that matrix. µ1, µ2, µ3, µ4 and µ represents mean for class1, class2, class3, class4 and entire data set. Mean is calculated by

$$\text{Mean}(\mu) = \frac{\sum \text{all values for each descriptor}}{\text{number of compounds}} \quad (1)$$

3. Mean corrected data for each class is calculated. x1o, x2o, x3o and x4o represents mean corrected data for each class. Mean corrected data is calculated by

Mean Corrected Data= data – mean for the entire set (2)

4. Co – Variance matrix for each class is calculated. c1, c2, c3 and c4 represent Co – Variance matrix for each class. Co - Variance matrix is calculated by

$$ci = \frac{(xio)^T * (xio)}{\text{number of compounds belonging to corresponding class}}$$

where i =1, 2, 3 and 4 (3)

For LDA algorithm, following steps is carried out in MATLAB.

5. Polled variance matrix is calculated, which is given by

$$c = \frac{1}{(\text{no of compounds})} \sum_{i=1}^{4} (ni * ci) \quad (4)$$

where ni – no of compounds in each class
ci – co – variance matrix for each call
i =1, 2, 3, 4

6. Inverse of the polled variance matrix is calculated. Inverse of the polled variance matrix is calculated by

cin = inv(c) (5)

where c - polled variance matrix

7. Prior probability vector p is found. If the prior probability is not known, it is assumed that it is equal to total sample of each group divided by the total sample.

$$Pi = \frac{ni}{N} \quad (6)$$

where ni – no of compounds in each class
N – total number of compounds
i=1, 2, 3, 4

8. Discriminant function is calculated for each compound in each class. The maximum dicriminant function represents the toxic class for each compound. Discriminant function is Clculated by

$$fi = (\mu i * cin * x(k)^T) - 0.5 * \mu i * cin * \mu i(k)^T + \ln(Pi) \quad (7)$$

where fi – discriminant function for each class
µi – mean value for each class
cin – inverse of polled variance matrix
x(k) – descriptors for each compound
i = 1, 2, 3 and 4
f1 indicates class1
f2 indicates class2
f3 indicates class3
f4 indicates class4

f1, f2, f3 and f4 are calculated for each compound. The highest of discriminant function indicates to which class the pesticide belongs.

For QDA algorithm, following steps is carried out in MATLAB.

9. Inverse of the Co – Variance matrix for each class is calculated. Inverse of the Co – Variance matrix is calculated by

   cini = inv(ci) (8)

   where i = 1,2,3 and 4

10. Prior probability vector p is found. If the prior probability is not known, it is assumed that it is equal to total sample of each group divided by the total sample.

$$Pi = \frac{ni}{N} \quad (9)$$

   where ni – no of compounds in each class
   N – total number of compounds
   i=1, 2, 3, 4

11. Discriminant function is calculated for each compound in each class. The maximum dicriminant function represents the toxic class for each compound. Discriminant function is calculated by

$$fi = (\mu i * cini * x(k)^T) - 0.5 * \mu i * cini * \mu i(k)^T + \ln(Pi) \quad (10)$$

   where fi – discriminant function for each class
   μi – mean value for each class
   cini – Inverse of the Co – Variance matrix for each class
   x(k) – descriptors for each compound
   i = 1, 2, 3 and 4
   f1 indicates class1
   f2 indicates class2
   f3 indicates class3
   f4 indicates class4

   f1, f2, f3 and f4 are calculated for each compound. The highest of discriminant function indicates to which class the pesticide belongs.

   LDA and QDA algorithm classified classes are compared with true class of each compound. Then success rate is calculated by

$$\text{success rate} = \frac{\text{no. of succesfull classification}}{\text{total no. of compounds}} \quad (11)$$

When the descriptors for new compound is given as input to the LDA or QDA algorithm, it calculates f1, f2, f3 and f4 and it tells to which class the compound belongs based on the highest value of f1, f2, f3 and f4.

Relative error1 is calculated by comparing true class and LDA or QDA classified class, where true class is differing by only one class more or less in LDA or QDA classified class.

For example if true class is 1 and LDA or QDA classified class 2 or vice versa, it corresponds to relative error1.

Relative error2 is calculated by comparing true class and LDA or QDA classified class, where true class is differing by only two classes more or less in LDA or QDA classified class.

For example if true class is 1 and LDA or QDA classified class 3 or vice versa, it corresponds to relative error2.

Relative error3 is calculated by comparing true class and LDA or QDA classified class, where true class is differing by three classes more or less in LDA or QDA classified class.

For example if true class is 1 and LDA or QDA classified class 4 or vice versa, it corresponds to relative error3.

Results

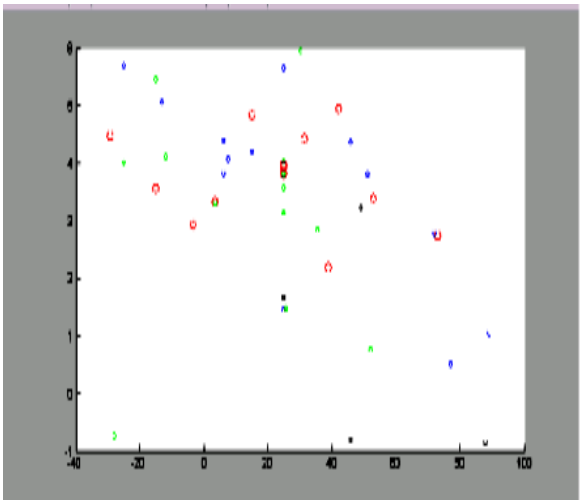

Figure 1. Input Data

Figure 1. Shows the plot of input data.
Red color circle indicates the values belonging to class1

Blue color circle indicates the values belonging to class2
Green color circle indicates the values belonging to class3
Black color circle indicates the values belonging to class4

True class, LDA classified class and QDA Classified class for each compound is as shown in table 3.

Table 3. True class, LDA classified class and QDA Classified class for each compound

Compound Name	True Class	LDA Classified Class	QDA Classified Class
Anilofos	1	1	4
Chloropyrifos	1	3	2
Isazofos	1	1	2
Phosalone	1	1	4
Prothiofos	1	3	2
Azamethiphos	1	1	1
Azinphos-methyl	1	2	1
Diazinon	1	1	1
Phosmet	1	3	1
Pirimiphos-ethyl	1	1	4
Pirimiphos- ethyl	1	1	1
Pyrazophos	1	3	1
Quinalphos	1	1	4
Azinphos-ethyl	2	2	1
Etrimfos	2	2	1
Fosthiazate	2	1	4
Methidathion	2	2	1
Piperophos	2	2	1
Triazophos	2	2	1
Dichlorvos	2	2	4
Disulfoton	2	2	4
Fenamiphos	2	3	2
Fenthion	2	3	2
Fonofos	2	3	1
Isofenphos	2	3	2
Methamidophos	2	1	4
Omethoate	2	2	1
Parathio	3	3	2
Parathion-ethyl	3	2	1
Phoxim	3	3	1
Sulfotep	3	3	1
Tribufos	3	4	2
Trichlorfon	3	3	2
Acephate	3	3	3
Dimethoate	3	3	2
Ethion	3	3	1
Ethoprop	3	3	1
Fenitrothion	3	2	1
Formothion	3	3	1
Phorate	3	2	2
Propetamphos	3	3	2
Sulprofos	4	3	1
Temefos	4	1	2
Terbufos	4	4	2
Thiometon	4	4	2

Success rate for LDA classification algorithm is calculated by comparing true class and LDA classified class for each compound. Eight descriptors are considered for each compound.

Success rate = 62.22 %

When the same descriptors are given as input to the LDA classification algorithm, it calculates f1, f2, f3 and f4 then predicts to which toxic class the new compound belong to, based on the highest value of discriminant function.

The descriptors for new compound is as shown

[42 4.96 1.12 2.03E-05 2.93E-06 9.17E-11 82 202]

LDA classification algorithm computes all discriminant functions and gives output as shown in table 4.

Table 4. Discriminant function for the new compound

f1	f2	f3	f4
15.8964	15.7196	15.4158	13.0663

The new compound analyzed belongs to class 1, since f1 is greater than other discriminant functions and f1 is discriminant function for class1.

Total number of elements versus maximum probable density is as shown in Fig. 2. This graph indicates the maximum discriminant function for each compound [8].

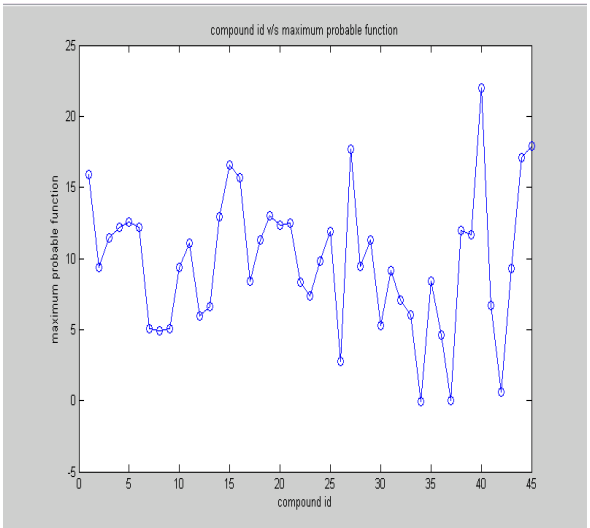

Figure 2: Total number of elements versus maximum probable density

Relative error1 for LDA algorithm is 70.58%. This error is in the range of acceptable error range, because LDA classified classes are only one class more or less as compared to true class.

Relative error2 for LDA algorithm is 23.52%. This error is not in the range of acceptable error range; because LDA classified classes are two classes more or less as compared to true class.

Relative error3 for LDA algorithm is 5.88%. This error is not in the range of acceptable error range; because LDA classified classes are three classes more or less as compared to true class. This error cannot be accepted in any case. For a highly toxic compound, LDA classifies as non toxic compound and for a non toxic compound, LDA classifies as highly toxic compound.

Success rate for QDA classification algorithm is calculated by comparing true class and QDA classified class for each compound. Eight descriptors are considered for each compound.

Success rate = 22.22%

When the same descriptors are given as input to the QDA classification algorithm, it calculates f1, f2, f3 and f4 then predicts to which toxic class the new compound belong to, based on the highest value of discriminant function.

The descriptors for new compound is as shown

[42 4.96 1.12 2.03E-05 2.93E-06 9.17E-11 82 202]

QDA classification algorithm computes all discriminant functions and gives output as shown in table 5.

Table 5. Discriminant function for the new compound

f1	f2	f3	f4
0.00	0.00	0.00	1.1237e+14

The new compound analyzed belongs to class 4, since f4 is greater than other discriminant functions and f4 is discriminant function for class4.

Relative error1 for QDA algorithm is 45.71%. This error is in the range of acceptable error range, because QDA classified classes are only one class more or less as compared to true class.

Relative error2 for QDA algorithm is 40%. This error is not in the range of acceptable error range; because QDA classified classes are two classes more or less as compared to true class.

Relative error3 for QDA algorithm is 14.28%. This error is not in the range of acceptable error range; because QDA classified classes are three classes more or less as compared to true class. This error cannot be accepted in any case. For a highly toxic compound, QDA classifies as non toxic compound and for a non toxic compound, QDA classifies as highly toxic compound.

Total number of elements versus maximum probable density is as shown in Fig. 3. This graph indicates the maximum discriminant function for each compound [8].

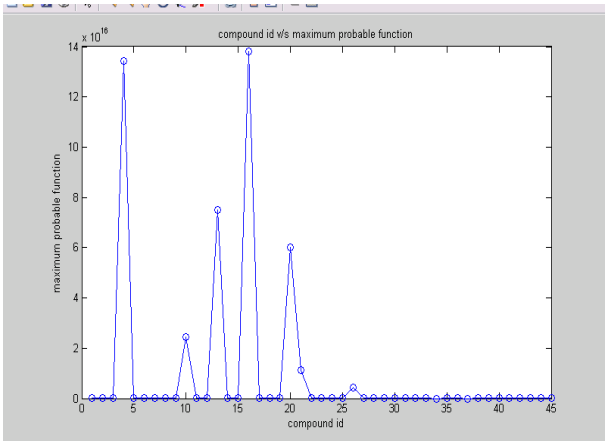

Figure 3: Total number of elements versus maximum probable density

The graph of total no of classes and number of classes which belongs to particular class is as shown in Fig. 4. This graph indicates the number of compounds belongs to each class.

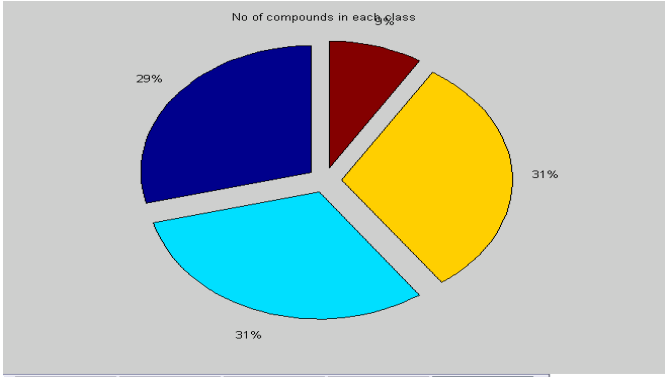

Figure 4: Total no of classes versus number of classes which belongs to particular class

Number of compounds which belongs to class1 is 13 out of 45, it is analyzed as 29% of total compounds, and this class indicates high toxicity level. In the figure 3, dark blue colour indicates the number of compounds belonging to class1.

Number of compounds which belongs to class2 is 14 out of 45, it is analyzed as 31% of total compounds, and this class indicates moderate toxicity level. In the figure 3, light blue colour indicates the number of compounds belonging to class2.

Number of compounds which belongs to class3 is 14 out of 45, it is analyzed as 31% of total compounds, and this class indicates low toxicity level. In the figure 3, yellow colour indicates the number of compounds belonging to class3.

Number of compounds which belongs to class4 is 4 out of 45, it is analyzed as 9% of total compounds, and this class indicates non-toxic level. In the figure 3, brown colour indicates the number of compounds belonging to class4.

Conclusion

Classification of the toxicity requires a high degree of experience from computational experts. In this research, LDA and QDA classification algorithms are used in toxicity class prediction and eight descriptors for each compound are considered. Success rate for LDA classification algorithm is found to be 62.22% and success rate for QDA classification algorithm is found to be 22.22%. If we consider more descriptors for compounds, success rate can be increased. Because some descriptor will help in predicting class correctly and some may not. QDA classification algorithm is not suitable for pesticide toxicity classification.

Future Scope

The toxicity level of a compound can be predicted by using different classification algorithms. It can be implemented using RDA (Regularized Discriminant Analysis), SIMCA (Soft Independent Modelling of Class Analogy), KNN (K Nearest Neighbors classification), and CART (Classification and Regression Tree) and many more. By comparing results of all the algorithms we can conclude which classification is suited for identification of toxicity class for compounds or pesticides.

Acknowledgments

The authors are thankful to IJIRTS Journal for the support to develop this document.

References

1. Emilio Benfenati1, Paolo Mazzatorta1, Daniel Neagu2, and Giuseppina Gini2, Combining classifiers of pesticides toxicity through a neuro-fuzzy approach, · Proceeding MCS '02 Proceedings of the Third International Workshop on Multiple Classifier Systems Pages 293-303 Springer-Verlag London,UK,2002
2. Giuseppina Gini, Emilio Benfenati, Daniel Boley, Clustering and Classification Techniques to Assess Aquatic Toxicity, **Volume:** 1, Fourth International Conference, 2000, 166 - 172
3. George C. J. Fernandez, Discriminant Analysis, A Powerful Classification Technique in Data Mining. http://www.lexjansen.com/wuss/2001/WUSS01036 .pdf
4. http://www.saedsayad.com/lda.htm
5. http://matlabdatamining.blogspot.in/2010/12/linear-discriminant-analysis-lda.html
6. U.S. National Library of Medicine, National Institutes of Health, TOXNET: Toxicology Data Network Fact Sheet, http://toxnet.nlm.nih.gov/
7. Kardi Teknomo,Linear discriminant ananlysis numerical example. http://people.revoledu.com/kardi/tutorial/LDA/
8. Dimitri P. Bertsekas and John N. Tsitsiklis, Introduction to Probability, 2nd ed. Athena Scientific, 2008

Biographies

JYOTHY S T obtained her BE in Electronics and Communication from SJMIT, Chitradurga in 2007. She has teaching experience of about 4 years. Currently she is studying in IV Semester M. Tech in Industrial Electronics at RNSIT, Bangalore. Her areas of interest are Microprocessors, HDL.

DEEPU KUMAR T L obtained his BE in Electronics and Communication from SJMIT, Chitradurga in 2000. He is having 12 years of experience in embedded system and RTOS. Currently he is working as a technical lead in Tata Elxsi Limited. His areas of interest are embedded system and RTOS.

Dr ANDHE PALLAVI obtained her BE in Instrumentation Technology in 1994, ME (ECE) in 1997 and PhD (EC) in 2007. She has teaching experience of about 18 years. She is currently working as Professor and Head of IT department, RNSIT, Bangalore. Areas of interest are DSP, Error Coding, Signal Processing, Networks. She is a life member of ISTE and ISOI societies. She has published many papers in national and international journals.