ISBN 978-81-934174-9-2
8th International Conference on Innovations in Engineering, Technology, Computers and Applied Sciences (IETCAS-2017)
Bangkok (Thailand) Dec. 25-26, 2017

A Framework to Study Data Mining to Predict Pesticide Uses

Masud Karim

University of Dhaka, Bangladesh
masud.mkarim@gmail.com

Abstract**—***Data mining shows an important role in decision making, future information prediction, data analysis, knowledge discovery etc. It plays vital role in business, medical, survey, social issues, agriculture etc. We can apply data mining theory on different sectors of agriculture. In this paper I have proposed a data mining framework to predict estimation usages of pesticide. In this framework data are collected from baseline and future information can be predicted by applying data mining classifier algorithm. I have analyzed performance of four classification algorithms: Naïve Bayes, kNN, J48 and Random Tree. I have applied these classifier algorithms on freely available pesticide dataset to predict uses of different pesticides. WEKA data mining tool is used to compare the results and it is found that the kNN classifier shows comparatively more accurate result.*

Keywords*: Knowledge Discovery, SVM, k Nearest Neighbor, Naive Bayes, Agriculture, Pesticide.*

1. Introduction

The major goal of this research is to predict the target class. Different techniques are used to predict and detect target classes from a large data source [1], [4], [5]. Very often it demands to make strategic policy, making corporate decision, market analysis, fraud detection etc from previous related data. Data mining techniques are used to predict required information and make decision from previously unknown information. It helps to find out hidden pattern from large data source. Mathematical tools for discovering patterns in large databases, along with stochastic modeling could contribute to better decision making in a range of fields. Researchers are doing data mining tasks in various fields from last of the years to predict, analysis and knowledge discovery [8].

2. Related Work

In [2] V.Thavavel and S.Sivakumar present text mining in distributed environment. They proposed a framework to analyze privacy preservation for distributed data mining. Unstructured data is converted into structured form using XML. Then they have applied their proposed method and data mining tools over the structured data.

In [7] Dr. S.Vijayarani uses a medical dataset for classification. The data set is collected from UCI repository to predict hard disease. They have analyzed three classification algorithms: logistics, multilayer perception and sequential minimal optimization. He has analyzed performance of these classification functions. WEKA data mining tool is used for comparative analysis. They have shown that logistics classification algorithm has the best result in this case. True positive (TP) rate, F Measure, ROC area and Kappa statistics are used to measure the accuracy.

In [9] Ashokkumar Vijaysinh Solanki applied open sourced data mining tool WEKA to predict sickle cell disease. He has used decision tree classifications. He also presented a comparison of two algorithms, J48 and Random Tree. After the experiments it was shown that using Random tree is better than J48. Random tree produces details decision tree comparing to J48, which is very much useful for further classification of each node. He has emphasized genetic disease, Sickle Cell Disease (SCD). This research is helpful to the society of medical sector and government department for the improvement of medical sectors.

In [12] Lambodar Jena and others used different classification algorithms in this research. They have applied these algorithms on chronic kidney disease related subjects. The data set is downloaded from UCI machine learning repository. They have used WEKA data mining tools for classification and describe a comparison analysis among the performance of six algorithms. Based on the analysis the researchers have illustrated that the multilayer perceptron has the best accuracy comparing Naïve Bayes, SVM, J48, conjunctive rules and decision tables. The objective of this research is to predict the target class accuracy for each case in the data. The researchers find out suitable algorithm for diagnosis and prediction of chronic kidney disease.

In [13] Aruna Govada and others have proposed an algorithm for classification. Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm is use to handle noisy data sets. Their proposed algorithm implements RIPPER at local level and then combines the output of local level into global level. In global level they have implemented distributed environment. They have used five distinct datasets distributed in different nodes. The accuracy of the algorithm is calculated by parameters, time taken for rule generation, accuracy in each iterations and testing accuracy.

In [15] Li Xiong and others work with private databases. They show a framework to analyze horizontally partion databases. They use k Nearest Neighbor for privacy preserving classifier which offers a trade-off between accuracy, efficiency and privacy through multi round protocols. They have presented some problems of using kNN regarding this research and discussed solutions. A comparison of privacy preserving and usual uses of kNN has been illustrated. Also the sensitivity is compared with the changes of various parameters. They have discussed results based on three different datasets.

In [17] Ms. Rupali Chikhale has discussed data mining algorithms, methods, techniques and their trends in using for distributed environment. Very often we use centralized data mining instead of distributed due to its good facilities likes fundamental rules of data mining algorithms, response time, lack of good uses of distributed resources etc. To solve this, the author emphasizes distributed application regarding resources in distributed environments and human factors. Some objectives of distributed data mining are also illustrated. The author has classified distributed data mining algorithms into three classes, based on multi agent system, meta learning and grid. All the types are discussed with their architecture. Their advantages, disadvantages and frameworks are also shown in comparative analysis.

3. Research Background

3.1. Pesticide in Agriculture

Pesticide is used to protect the plant from insects. These are chemical or biological agents that deter, incapacitates, kill, or discourage pests. Although it saves the plants but its uses sometimes damage environment and become harmful for food. Many active ingredients are used in pesticides which have some bad effects. So its proper uses are very important.

3.2. The Dataset

The data set is collected from [18] which is a survey data of United States Department of Agriculture Crop Reporting Districts. It includes country level pesticide using estimation of pesticides used for crops growth in the countries of United States.

TABLE I. A Sample of Dataset

Compound	Year	State Code	County Code	Low Estimate	High Estimate
2,4-D	2014	1	1	1698.6	1885.5
2,4-D	2014	1	3	7513.6	8472.4
Acetochlor	2014	5	17	638	25512.6
Acetochlor	2014	5	19	2.25	3.8
Bacillus Firmus	2014	31	175	90.8	90.8
Metalaxyl	2014	31	137	78.9	79.2
Metalaxyl	2014	31	139	87.4	87.4
Pendimethalin	2014	53	37	182.3	246.8
Pendimethalin	2014	53	39	514.4	740.2
Phosphoric Acid	2014	8	23	3.65	5095.7
Phosphoric Acid	2014	8	29	17.4	17.4
Sulfentrazone	2014	51	183	28.4	37.7
Trifloxystrobin	2014	38	43	288	250.5
Trifloxystrobin	2014	38	45	192	573.4

3.3. Classification and Classification Algorithms

The power of data mining is physically shown by data mining algorithms. These algorithms create data mining model from data. The model is created by analyzes of the data, looking for specific types of patterns or trends. From this patterns or trends there are results. The algorithms also use the results of this analysis to define the optimal parameters for creating the mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics. Then we get the actionable knowledge [1], [5], [12]. There are many algorithms developed for classification. Different algorithm has different uses. Choosing the right algorithm is also a challenging task. Different algorithms have different objectives to perform different business tasks [3], [4], [13]. Sometimes algorithms are in same style but different in presentation of output results. It is also a difficult task to classify the data mining algorithms in specific fashion.

4. Proposed Models

Due to increasing data and flow of information data mining algorithms are becoming popular to classify and knowledge discovery [4], [5], [8], [11]. Data mining framework is being using to discover knowledge [11], [14]. The model in figure 1 shows steps to create such type of framework. This model predicts future data from baseline database.

Objective of this research is to build a common framework that can be used in future. From the baseline survey data are collected and stored in database. We need to change some of this collected data. Some data or information may not be required to build the framework. It is the pre-process by which data are being made suitable for next procedures. It may require preprocess the data stored in database. Whenever collecting the data there are many information and all the information is not necessary for applying data mining tools. Sometimes we need to convert data into other format or change the data type. After the preprocessing data, we will apply data mining tools & techniques to build different models. Data mining tools are applied on these data. Then models are built to predict the target group. Among these models the best model is selected for further use. The selection of model depends on accuracy, time taken to build the model, classification correctness and others performance of data mining. The figure 1 shows a typical framework to select the target group.

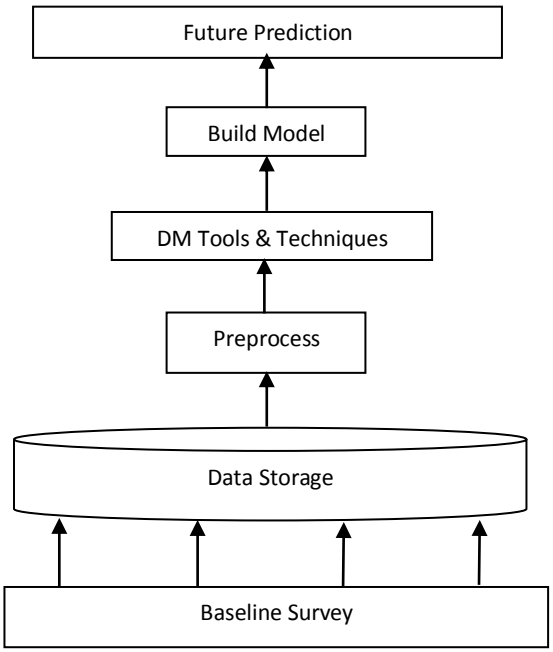

Fig 1. A Typical Diagram of Proposed model.

5. Experiment Setup

To predict the uses of pesticide, I have applied classifier algorithms. Different powerful classifier algorithms are being used by data mining researchers [3], [10]. I have used four algorithms to predict pesticide uses in agriculture. The algorithms are used as defined in WEKA [2], [6], [7], [12]. Some researchers use a combination technique with classification algorithms [1], [4], [5], [13], [16]. Performances of these algorithms are also compared to find the best classifier. Two publicly available datasets [18] are used to train and test the model. It contains pesticide using data of different countries in United State. The first dataset is collection of data in the year

2014 and another of 2015. Dataset of 2014 contains 392433 records and 2015 contains 367714 records. A total of 388 pesticides used in 2014 and 343 pesticides in 2015 dataset. The datasets are therefore not a small dataset. The datasets are little bit modified for selecting the features [14].

To formalize the data and usable for this research I have added another attribute: 'status' of pesticide uses. Status includes uses of pesticides in high (H), low (L), not uses (N) and equal uses (E) of high and low estimation. Status depends on pesticide uses in high estimation and low estimation columns. After that ARFF file is created from the dataset. ARFF is Attribute Relation File Format is an ASCII text file. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for using with the WEKA machine learning software. This file describes a list of instances sharing a set of attributes.

TABLE II. A sample of filtered dataset

Compound	State Code	County Code	Status
2.4-D	1	1	H
2.4-D	1	3	H
Acephate	51	57	E
Acephate	51	59	H
Bacillus-Pumilis	41	21	N
Bacillus-Pumilis	41	27	H
Captan	12	83	E
Captan	12	86	E
Fluroxypyr	41	5	H
Fluroxypyr	41	9	H
Metconazole	55	93	H
Metconazole	55	95	H
Phosmet	12	71	E
Phosmet	12	75	H
Picloram	51	163	H
Picloram	51	165	E

TABLE III. Dataset Description

Attributes	Descriptions
Compound	Nominal value, These are the name of pesticides.
State Code	Numeric code to identify state.
County Code	Numeric code to identify country.
Status	Nominal value for target class.

6. Experimental Result & Discussion

In this experiment I have used four classifier algorithms: Naïve Bayes, K nearest neighbor, J48 and random tree. The collected data is preprocessed for running the algorithm. After the training and testing, performance are found as the below.

TABLE IV. Classifier Performance

Dataset	Instances	Classifier	Correctly Classified (%)	Incorrectly Classified (%)
2014	133427	Naïve Bayes	88.49	11.5
		kNN	91.51	8.48
		J48	97.8	2.19
		Random Tree	90	9.99
2015	125023	Naïve Bayes	89.39	10.6
		kNN	92.88	7.11
		J48	90.03	9.96
		Random Tree	90.03	9.96

It seems that for dataset 2014 the J48 classifies more correctly but it takes more time to build the model. kNN is second position to be classified more instances and it takes less time than J48. For 2015 dataset kNN correctly classifies more instances but Naïve Bayes takes less time than kNN.

TABLE V: Accuracy of Classifiers.

DATASET	Classifier	Accuracy (%)	Time to Build Model (Second)
2014	Naïve Bayes	88	0.24
	kNN	91	0.06
	J48	98	50.27
	Random Tree	90	5.15
2015	Naïve Bayes	89	0.21
	kNN	93	0.25
	J48	90	5.66
	Random Tree	90	4.59

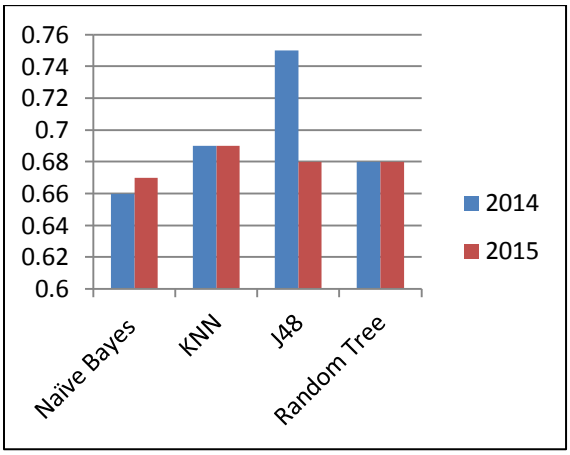

Fig 2. Accuracy of Classifiers

In terms of accuracy and time taken to build the model in an average of two dataset, kNN shows more accurate result. Though, the J48 shows more accuracy for the dataset 2014.

TABLE VI: Detail Accuracy of Classifier

Dataset	Classifier	Precision	Recall	F Measure	ROC Area
2014	Naïve Bayes	0.64	0.66	0.638	0.748
	kNN	0.69	0.69	0.69	0.73
	J48	0.74	0.75	0.74	0.83
	Random Tree	0.67	0.68	0.68	0.70
2015	Naïve Bayes	0.65	0.67	0.64	0.74
	kNN	0.69	0.69	0.69	0.72
	J48	0.68	0.68	0.68	0.699
	Random Tree	0.68	0.68	0.68	0.699

From the above table and classification results it is understood that kNN and J48 bring better result. So for shortage of page limit I have illustrated ROC curve and error curve of these two classifiers. ROC curve is shown for predicting pesticide used in high level and low level.

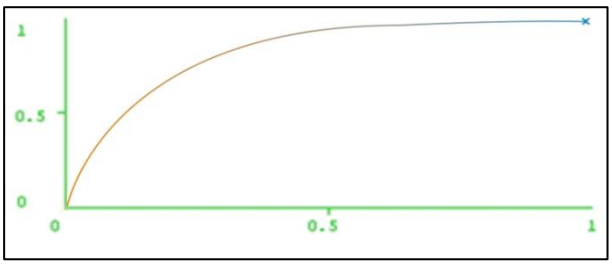

Fig 3. ROC curve of J48 Classifier Using High uses of Pesticide in 2014.

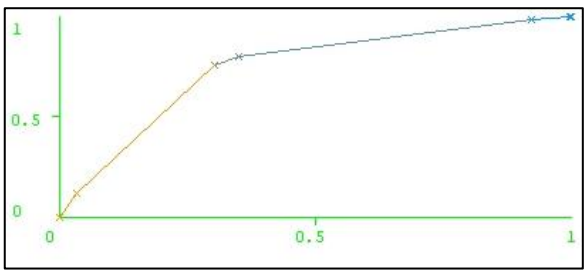

Fig 4. ROC curve of kNN Classifier Using High uses of Pesticide in 2014.

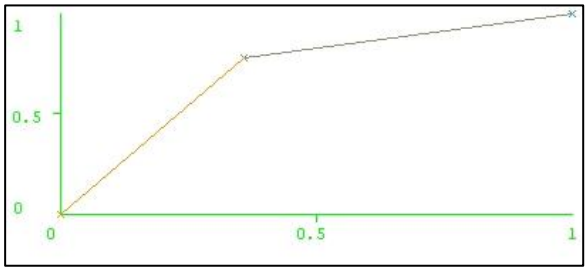

Fig 5. ROC curve of J48 Classifier Using High uses of Pesticide in 2015.

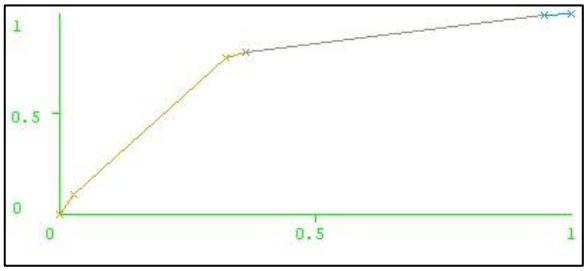

Fig 6. ROC curve of kNN Classifier Using High uses of Pesticide in 2015.

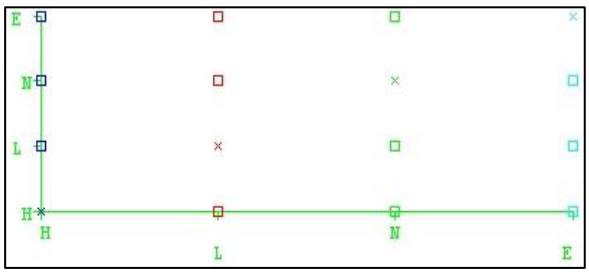

Fig 7. Visual Error Curve of J48 Using data set of 2014

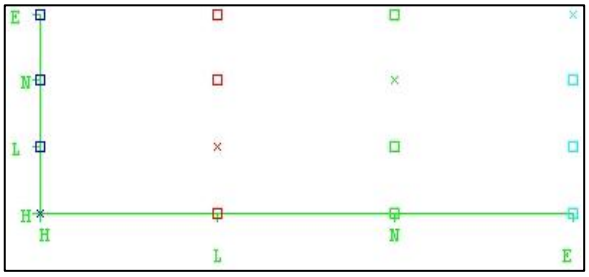

Fig 8. Visual Error Curve of kNN Using data set of 2015

7. Conclusion

In this study I have proposed a model to predict a target class and analyzed performance of four classifier algorithms. These algorithms are used to build the model. Freely available pesticide dataset of the year 2014 and 2015 are used in this research as test basis. A classification model of pesticide uses are shown here. This proposed model can be used to predict future information. For further development the model can be improved for using in distributed fashion of real time data.

8. Acknowledgement

I acknowledged ICT Division, Ministry of Posts, Telecommunications and IT, Government of the People's Republic of Bangladesh for supporting. I thank to students of Daffodil Institute of Information Technology for experimental task and also department of Computer Science and Engineering, University of Dhaka.

9. References

[1] Daranee Thitiprayoonwongse, Prapat Suriyaphol, Nuanwan Soonthornphisaj, “A Data Mining Framework for Building Dengue Infection Disease Model”, The 26th Annual Conference of the Japanese Society for Artificial Intelligence, 2012.

[2] V.Thavavel and S.Sivakumar, “A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment”, International Journal of Computer Science Issues, Vol 9, Issue 1, No 2, January 2012, ISSN (Online): 1694-0814.

[3] D.Sindhuja, “A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder”, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016, pg. 483-488.

[4] S. V. S. GANGA DEVI, “A Survey on Distributed Data Mining and Its Trends”, International Journal of Research in Engineering & Technology (IMPACT: IJRET), ISSN (E): 2321-8843; ISSN (P): 2347-4599, Vol. 2, Issue 3, Mar 2014, 107-120.

[5] A. B. Devale and Dr. R. V. Kulkarni, “Applications of Data Mining Techniques In Life Insurance”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.4, July 2012. DOI: 10.5121/ijdkp.2012.2404

[6] Prathmesh Kulkarni, “Big Data Mining Tools”, International Journal of Scientific Development and Research (IJSDR), April 2016, ISSN: 2455-2631.

[7] Dr. S.Vijayarani, S.Sudha, “Comparative Analysis of Classification Function Techniques for Heart Disease Prediction”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 3, May 2013, ISSN (Print) : 2320 – 9798, ISSN (Online): 2320 – 9801.

[8] Shraddha Masih, Sanjay Tanwani, "Data Mining Techniques in Parallel and Distributed Environment- A Comprehensive Survey", International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 3, March 2014.

[9] Ashokkumar Vijaysinh Solanki, "Data Mining Techniques Using WEKA classification for Sickle Cell Disease", International Journal of Computer Science and Information Technologies Vol.5 (4), 2014, 5857-5860.

[10] N. Mlambo, "Data Mining: Techniques, Key Challenges and Approaches for Improvement", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 3, March 2016, ISSN: 2277 128X.

[11] Mario Cannataro, Andrea Pugliese, "Distributed Data Mining on Grids: Services, Tools, and Applications", IEEE Transactions on Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 34, No. 6, December 2004.

[12] Lambodar Jena, Narendra Ku. Kamila "Distributed Data Mining Classification Algorithms for Prediction of Chronic-Kidney-Disease", International Journal of Emerging Research in Management &Technology, ISSN: 2278-9359 (Volume-4, Issue-11), November 2015.

[13] Aruna Govada, Varsha S. Thomas, Ipsita Samal, Sanjay K. Sahay, "Distributed multi-class rule based classification using RIPPER", 2016 IEEE International Conference on Computer and Information Technology, DOI 10.1109/CIT.2016.111.

[14] Jun Wang, Jin-Mao Wei, Zhenglu Yang and Shu-Qin Wang, "Feature Selection by Maximizing Independent Classification Information", IEEE Transactions On Knowledge And Data Engineering,2016. DOI 10.1109/TKDE.2017.2650906.

[15] Li Xiong, Subramanyam Chitti, Ling Liu, "Mining Multiple Private Databases Using a kNN Classifier", SAC'07 March 11-15, 2007, Seoul, Korea. Copyright 2007 ACM 1-59593-480-4/07/0003.

[16] Heling Jiang, An Yang, Fengyun Yan, Hong Miao, "Research on Pattern Analysis and Data Classification Methodology for Data Mining and Knowledge Discovery", International Journal of Hybrid Information Technology, Vol.9, No.3 (2016), pp. 179-188

[17] Ms. Rupali Chikhale, "Study of Distributed Data Mining Algorithm and Trends", National Conference on Recent Trends in Computer Science and Information Technology (NCRTCSIT-2016), IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661,p-ISSN: 2278-8727, PP 41-47

[18] https://www.kaggle.com/usgs/pesticide-use