Improving Classification Accuracy in Using Evolutionary Decision Tree Filtering in Big Datasets

Nikhat Raza Khan and Ashish Jain
IES College of Technology Bhopal, India

Keywords: Large Datasets, Erroneous Data, Evolutionary Decision Tree, Classification Accuracy, Genetic Algorithm.

Abstract: Large datasets frequently have data value errors, which can be a serious issue. The method described in this paper uses a classification filter to locate and eliminate incorrect training instances from large datasets. The suggested approach uses an evolutionary decision tree technique as a filter classifier and aims to increase classification accuracy by pursuing a broad search in the problem space using a genetic algorithm. This method improves classification accuracy compared to conventional decision tree algorithms by using classification accuracy as a fitness function instead of local greedy search on data sets. The proposed approach performed better than earlier methods in terms of decision tree size and classification accuracy after being tested on the UCI repository data set. The results demonstrate the effectiveness of the proposed method in handling datasets with errors in data values.

1 INTRODUCTION

Data mining is the process of extracting knowledge from large amounts of data available from various sources. The knowledge discovery process is comprised of the following iterative subtasks (Han and Kamber 2006): data cleaning to remove noise and inconsistency in data, data integration to combine multiple data sources, data selection, data transformation, generating data patterns, pattern evaluation, and knowledge representation. Lavarac (Lavarac et al 2018) proposed avoiding the influence of erroneous data on hypothesis by removing it from training data before induction. In a data set, an outliner is an instance that is significantly divergent or inconsistent from the rest of the data (Xiong et al 2020). This paper proposed an evolutionary tree as a classification filter, and by using a genetic algorithm, the filter performance is extremely good, providing noiseless data to the classifier, resulting in very high classification accuracy and smaller sized trees on the classifier.

2 RELATED WORK

The problem of noise elimination is critical. Several researchers have proposed various approaches to data cleaning (Xiong et al 2020 – Loureiro 2021). This includes the use of MDL. Principle, neural networks, filters, Occam's razor, and a few other techniques. Different authors employ genetic algorithms to provide a global search through space in multiple directions at the same time. To evolve optimal decision trees, various authors have proposed using methodologies that integrate genetic algorithms and decision tree learning. Despite the fact that the methods differ, the goal is to obtain optimal decision trees. GATree, a genetically evolved decision tree, was proposed by A. Papagelis and D. Kalles in 2021. The genetic algorithm is used to evolve binary decision trees directly. GALib's tree representation was used to generate a population of binary decision trees.

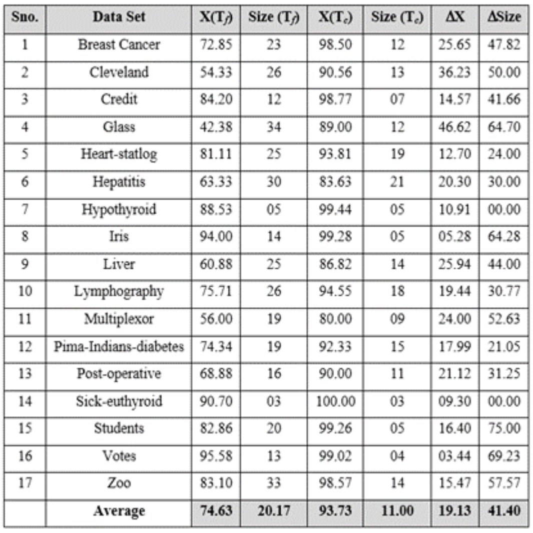
3 PROBLEM DESCRIPTION AND PROPOSED ALGORITHM

Let $\mathbf{T_f}$ represent a collection of all available **n** training instances. Let **t** represent the training instance. When hypothesis **H** is induced on $\mathbf{T_f}$, let $\mathbf{T_c}$ represent the set of instances that follow the hypothesis **H** and let $\mathbf{T_e}$ represent the set of instances that do not follow the hypothesis **H**. Thus, Hypothesis **H** is referred to as exceptional data, and $\mathbf{T_c}$ is referred to as clean data.

Khan, N. and Jain, A.
Improving Classification Accuracy in Using Evolutionary Decision Tree Filtering in Big Datasets.
DOI: 10.5220/0012610000003739

In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 189-191
ISBN: 978-989-758-661-3

Table 1: Results from Evolutionary Decision Tree Algorithm.



Sno.	Data Set	$X(T_f)$	Size (T_f)	$X(T_c)$	Size (T_c)	ΔX	ΔSize
1	Breast Cancer	72.85	23	98.50	12	25.65	47.82
2	Cleveland	54.33	26	90.56	13	36.23	50.00
3	Credit	84.20	12	98.77	07	14.57	41.66
4	Glass	42.38	34	89.00	12	46.62	64.70
5	Heart-statlog	81.11	25	93.81	19	12.70	24.00
6	Hepatitis	63.33	30	83.63	21	20.30	30.00
7	Hypothyroid	88.53	05	99.44	05	10.91	00.00
8	Iris	94.00	14	99.28	05	05.28	64.28
9	Liver	60.88	25	86.82	14	25.94	44.00
10	Lymphography	75.71	26	94.55	18	19.44	30.77
11	Multiplexor	56.00	19	80.00	09	24.00	52.63
12	Pima-Indians-diabetes	74.34	19	92.33	15	17.99	21.05
13	Post-operative	68.88	16	90.00	11	21.12	31.25
14	Sick-euthyroid	90.70	03	100.00	03	09.30	00.00
15	Students	82.86	20	99.26	05	16.40	75.00
16	Votes	95.58	13	99.02	04	03.44	69.23
17	Zoo	83.10	33	98.57	14	15.47	57.57
	Average	**74.63**	**20.17**	**93.73**	**11.00**	**19.13**	**41.40**

Let | T_f | be the training data set's cardinality. So

$$T_f = T_c \cup T_e .$$

Let Classification accuracy on test data instance be **X(t)**. Now ∀ **t** on **H**,

if **X(t) = 1** ⇒ **t** ∈ T_c and if **X(t)** = 0 ⇒ t ∈ T_e.

The proposed evolutionary decision tree algorithm is robust and efficient. It is used as a classification filter function as follows:

```
1. Induce Hypothesis H on $T_f$ with Evolutionary Decision Tree
2. for every Instance t ∈ $T_f$ , Do:
3.     if H correctly classifies t, then
               t ∈ $T_c$
4.     else
               t ∈ $T_c$
   end for
5. Induce Hypothesis $H_c$ on $T_c$ with Evolutionary Decision Tree as Final Classifier
6. End
```

G.H. John in 2021, demonstrated that locally uninformative or harmful instances are globally uninformative. John George's method removes a misclassified instance from training data by running the algorithm multiple times as many times as the number of outliners, whereas the proposed method removes outliner data from training data in a single run of the algorithm.

4 METHOD OF EXPERIMENTATION

T_f was tested for outliners using 18 data sets from the University of California Irvine repository (Neuman et al 2020). For experiments, the GATree algorithm (Papagelis and Kalles 2021) was modified in this implementation.

As a result, there were two data sets, T_c which contain clean data and T_e which contain outliner data. The abbreviations for a classifier's classification accuracy on a full training set and a cleaned data set are **X(T_f)** and **X(T_c)**. The size (T_f) and size (T_c) are abbreviations for the size of decision trees on hypothesis on **T** and **T** as training data, respectively.

The percentage difference in the tree size (**Δ** Size) and the absolute difference in accuracy (**Δ** Accuracy) between the trees built on full training instances and the trees built on cleaned training instances.

The following equations explain the Tree size (**Δ**Size) and the absolute difference in accuracy (**Δ**Accuracy)

$$\mathbf{\Delta Size} = \frac{size(T_f) - size(T_c)}{size(T_f\ )} \mathbf{* 100} \qquad (1)$$

$$\mathbf{\Delta Accuracy = X\ (T_c) - X\ (T_f)} \qquad (2)$$

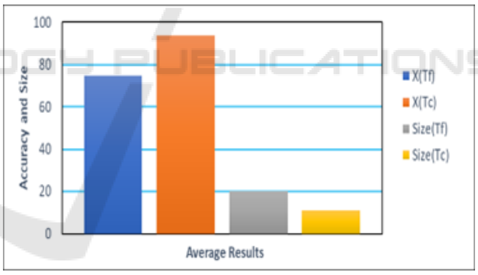


Figure 1: Comparison of Average Accuracy and Size.

5 RESULT

Table 1 summarizes the results of the evolutionary decision tree algorithm using the experimental method described above. The table below summarizes the percentage decrease in Tree size (**Δ**Size) and the absolute difference in accuracy (**Δ**Accuracy) between trees built on full training instances and trees built on cleaned training instances. The results for a robust and efficient evolutionary decision tree are also shown in Table 1. The Size column of the table comparison displays the percentage reduction in tree size between trees built on full training instances and trees built on cleaned training instances. It has been tested on full training instances, and the trees have been built on cleaned training instances.

The average absolute observed percentage reduction in tree size is significant, i.e. a 41.40% reduction in tree size. The accuracy column displays the absolute difference in accuracy between trees built. The average difference in accuracy for a cleaned data set is 19.13%. When the accuracies are compared, we can argue that evolutionary trees with cleaned data sets are more accurate. The proposed work has an average classification accuracy of 93.73%, compared to an average of **74.63**% with outliners data. Thus, we can argue that robust efficient evolutionary trees improve classifier performance.

6 CONCLUSION

Based on a statistical approach, this paper proposed a technique for dealing with outliers in data. When the outliners are removed, the induced patterns become more accurate and extremely simple. The results we obtained validate the use of the proposed techniques for this task. Furthermore, when compared to previous approaches to the same data, the results clearly outperform them, even with the same level of erroneous data. The proposed algorithm employs an evolutionary decision tree as a filter classifier for training data in order to pursue a global search in the problem space with classification accuracy as a fitness function while avoiding a local optimum, and the final classifier employs a cleaned data set. As a result of this combination of techniques, we have a robust and efficient classifier.

REFERENCES

J. Han, M. Kamber, (2006), *Data Mining: Concepts and Techniques* (2nd edition), Morgan Kaufman Publishers.

N Lavarac, D. Gamberger, P. Turney (2018), Cost-sensitive feature reduction applied to a hybrid genetic algorithm, in 7th *International workshop on algorithmic learning theory* (ALT'18), Sydney, Australia, October 2018, pg.127-pg.134

H. Xiong, G. Pande, M. Steinbach, V. Kumar (2020), Enhancing data analysis with noise removal, *IEEE Transaction on Data Engineering, Vol. 37*(issue 3)

D. Gamberger, N. Lavrac, S. Dzeroski (2022), Noise Elimination in Inductive Concept Learning: A Case Study in Medical Diagnosis, *International workshop on algorithmic learning theory*, Sydney, Australia

A. Aming, R. Agrawal, P. Raghavan (2019), *A linear method for deviation detection in large databases in KDDM*, pg.164-pg.169

Guyon, N. Matic, V. Vapnik (2019), Discovering informative patterns and data cleaning, *Advances in knowledge discovery and data mining*, pg.181-pg.203

D. Gamberger, N. Lavrae (2021), Conditions for occam's razor applicability and noise elimination, *European conference on machine learning (Springer)*, pg.108-pg.123

E. M. Knorr, R. T. Ng (2017), A unified notion of outliers: properties and computation, *3rd International conference on knowledge discovery and data mining*

E. M. Knorr, R. T. Ng (2018), *Algorithms for mining distance-based outliers in large datasets*, pg.392-pg.403

D. Tax, R. Duin. (2021), Outliner detection using classifier instability, *Workshop on statistical pattern recognition*, Sydney, Australia

C. Brodley, M. Friedl (2020), *Identifying Mislabeled Training Data*, JAIR, pg.131-pg.161

D. Gamberger, N. Lavrac, C. Groselj (2022), *Experiments with Noise Filtering in a Medical Domain, in ICML*, Morgan Kaufman, San Francisco, CA, pg.143-pg.51

S. Schwarm, S. Wolfman (2020), Cleaning Data with Bayesian Methods, *Final Project Report for University of Washington Computer Science and Engineering*, CSES74

S. Ramaswam, R. Rastogi, K. Shim (2022), *Efficient Algorithms for Mining Outliers from Large Data Sets*, in ACM SIGMOD, Vol. 29, pg.427-pg.438

V. Raman, J.M. Hellerstein (2020), *An Interactive Framework fo*
r Data Transformation and Cleaning, Technical Report University of California, Berkeley

J. Kubica, A. Moore (2019), Probabilistic Noise Identification and Data Cleaning, *IEEE International Conference on Data Mining*

V. Verbaeten, A. V. Assche (2020), Ensemble Methods for Noise Elimination in Classification Problems, *Multiple Classifier Systems*, (Springer)

J. A. Loureiro, L. Torgo, C. Soares, (2021), Outlier Detection Using Clustering Methods: A Data cleaning application, *Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector*, Bonn, Germany.

A. Papagelis, D. Kalles (2021), GATree: Genetically Evolved Decision Trees, *International Conference on Tools with Artificial Intelligence*, pg.203-pg.206

G. H. John (2021), *Robust Decision Trees: Removing Outliers from Databases*, 1st ICKDDM, pg.174-pg179.

D. J. Newman, S. Hettich, C. L. Blake, C. J. Merz (2020), *UCI Repository of Machine Learning Databases, Department of Information and Computer Science, University of California*, Irvine