International Journal of Database Theory and Application
Vol.9, No.3 (2016), pp.87-94
http://dx.doi.org/10.14257/ijdta.2016.9.3.10

An Optimized Decision Trees Approach for Knowledge Discovery Using Orthogonal Radom Matrix Projection with Outlier Detection

Mohammed Moulana* and Mohammed Ali Hussain**

**Research Scholar, Dept. of Computer Science, Bharathiar University, India.*
***Professor, Dept. of Electronics and Computer Engineering, KLEF University, India.*
mdmmsc@gmail.com, alihussain.phd@gmail.com

Abstract

In data mining one of the challenging problems is how to handle high dimensional and complex datasets. Decision trees when applied to high dimensional and complex datasets produce decision trees which are very complex in nature and thereby reducing generalization. To address this issue we propose an algorithm know as Radom Matrix Projection with Outlier Detection (RMPOD). The proposed algorithm is validated on 24 UCI datasets against accuracy and tree size metrics. The results of the proposed algorithm with compared algorithm suggest an improvement in accuracy and tree size for better generalization.

***Keywords**: Classification, decision trees, random projection, outlier detection, RMPOD*

1. Introduction

In data mining is the process of knowledge discovery is done from the hidden data sources. In knowledge discovery there may different techniques; Classification is the process of predicting the class of an unknown instances. In clustering the set of instances are grouped in some class by analyzing the characteristics and properties of the instances. An association analysis is the process of finding novel associations or patters which occur frequently in the data source. In classification the unknown or new instances are classified into the predefined classes. The process of classification can be done by using different models such as decision trees, neural networks and support vector machines *etc.* One of the benchmark and popular decision tree model is C4.5.

The process of classification in data mining is also known as supervised learning since a specific training phase is used to build the model before performing testing. Classification is used for decision making and decision making can be considered as one of the characteristics of intelligence. Generally decision tree models are designed to maximize the accuracy and minimize the tree size. Decision trees efficiency will drop down when handling with datasets of complex nature. A novel approach for decision trees are needed to handle specific issue of unique and abnormal datasets.

Researchers have applied statistical approaches for text and image data analysis [1] and suggested in future work “A still more realistic application of random projection would be to use it in a data mining problem”. Therefore, as an extension of the research we address different possibilities for applicability of random transformation techniques on decision trees for different high dimensional and complex data sources. Another objective of this research study is to fine tune and apply the best random transformation technique for real world datasets.

ISSN: 2005-4270 IJDTA

2. Literature Review

The area of decision tree learning has a vast data of recent and premier publications. In the below part of the manuscript we presented recent approaches adopted by differ researchers for decision tree exploration.

In [2] authors analyzed patters of diseased data using classification algorithms of decision tree and naive bayes. In [3] authors have analyzed different approaches for efficient prediction of credit default using attributes selection and using C4.5 as base algorithm. The details of patients regarding the ECG signals are classified into different predefined classes using a efficient decision tree approach [4].

In [5] authors have proposed an efficient classifier which forecasts the future occurrence of the incidences using decision tree approach. In [6] authors have proposed a decision tree approach for fraud detection in complex datasets. In [7] authors have presented Local Outlier Factor (LOF) technique to identify diverse trees in random forest for exhibiting superior accuracy.

In [8] authors have proposed model for addressing the challenges for imperfect data with naive bayes and logistic regression as the base learners. In [9] authors have proposed algorithms for distributed data mining for intensifying performance. In [10] authors have discussed classification techniques for different tree based approaches for novel class detection in evolving data stream. The recent literature review of decision trees suggests the need of improved classification algorithms for varied data sources.

3. Radom Matrix Projection with Outlier Detection (RMPOD)

The proposed Radom Matrix Projection with Outlier Detection (RMPOD) algorithm works with the simple principle of linear attribute transformation for optimization. The applicability transformation can retain the real representation of the data source in spite of the reduction of the data. The following are the conditions for better applicability of random projections:

1. Due to high dimensionality the computation of principal component is too expensive.

2. In data streams all the data can't be accessed at once.

3. The dimensionality of the data is low but it is not projected in near the linear Subspace.

In the proposed algorithm, a random matrix projection in specific orthogonal direction is implemented. The acceptable orthogonal direction for efficient data transformation is initially identified. In the later stages the following three main operations yield to the optimization: Conversion of attribute data using orthogonal random Projection, Outlier Detection and decision tree induction.

Random Projections:

Random Projections is one of the best techniques for handling complex data sources. The direction in which the random projection is done is independent of the data. In the proposed method transformation is done in such a way that the originality of the data is preserved in spite of the projecting the data in the random space using the principle of Euclidean distance [1].

In this experimental implementation both sparse and Gaussian entities with random matrix are used. The following are the main characteristics of the random space that help in the optimization of the transformation process: the length of the rows is equal and they are orthogonal in nature to each other are only considered for the process of random matrix projection. The above conditions are implemented in the process of matrix projection to achieve approximate isometry. Thus, in practical applicability there is no necessity of orthogonalise the random projection matrix after transformation [11].

Orthogonal Projections:

Let us say we have an orthonormal basis for a linear subspace, stacked into a matrix: Q = [q1 q2 : : : q`]. Then QQ^T is a projection matrix which operates on any matrix A to project it orthogonally onto the subspace spanned by Q, which we denote $P_Q(A)$.

$P_Q(A) = QQ^T A$:

The random matrix is multiplied in the process of orthogonal projections to make the result uncorrelated. The dimension size computation of random matrix projection is minimized for a particular dimension size. In our experimental simulation two categories are random projections are considered. They are as follows:
The algorithm for RMPOD is given below,

Algorithm: Radom Matrix Projection with Outlier Detection (RMPOD)

Input: D – Data Partition, A – Attribute List,*m x n* matrix A, number of samples S=k+p.

Output: A'=R *m x n*, A Decision Tree (D, A'')

Procedure:

Attribute Transformation (D, A)

```
Draw a Random Test matrix Ω n x s.
From the product Y m x s = A Ω.
Compute a orthonormal basis Q m x k for the rage of Y via SVD.
return A'=$QQ^T$A.
return (D, A')
Outlier Detection (D, A'')
return (D', A'')
Create a node N
If samples in N are of same class, C then
    return N as a leaf node and mark class C;
If A' is empty then
    returnN as a leaf node and mark with majority class;
else
    apply Gain Ratio(D', A')
    label root node N as f(A')
    for each outcome j of f(A') do
        subtreej =New Decision Tree(Dj',A')
        connect the root node N to subtree j
    endfor
endif
endif
Return N
```

Sparse: In sparse matrix projection different probabilities of 1/6 and 2/3 are considered [12, 13]. The discrete distribution over the values is done in one of the simplest sparse distribution. The equation used here is,

sqrt(3) * { -1 with prob(1/6),
0 with prob(2/3),
+1 with prob(1/6) }

Gaussian: The standard normal variates are considered for one of the Gaussian choices. In the best characteristics of Gaussian distribution is the generation of dense matrix projection. In the case of low dimensional space the geometry of the random matrix projection is preserved.

Decision tree induction:

The improved performance or fine tuning can be achieved in the final stage by eliminating the noisy and outlier instances from the data source. The improved data source is used for inducing decision trees. The base algorithm used for induction of decision tree is C4.5.

4. Experimental Design and Algorithms Compared

In this experimental setup, the researchers have used 24 UCI [14] data sets which are publicly available. The details such as number of instances, missing values, number of numeric and nominal attributes are given in the Table 1.

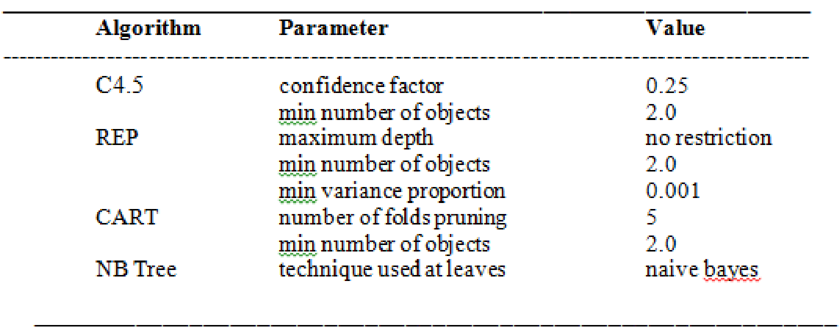
Table 1. The 24 UCI Datasets and their Properties

S.no.	Dataset	Instances	Missing values	Numeric attributes	Nominal attributes	Classes
1.	Anneal.ORIG	898	Yes	5	28	6
2.	Balance-scale	625	No	4	0	3
3.	Breast-cancer	286	Yes	0	9	2
4.	Breast-w	699	Yes	9	0	2
5.	Horse-colic	368	Yes	7	15	2
6.	Credit-a	690	Yes	6	9	2
7.	Credit-g	1,000	No	7	13	2
8.	Pima diabetes	768	No	8	0	2
9.	Glass	214	No	9	0	6
10.	Heart-c	303	Yes	6	7	2
11.	Heart-h	294	Yes	6	7	2
12.	Heart-statlog	270	No	13	0	2
13.	Hepatitis	155	Yes	6	13	12
14.	Ionosphere	351	No	34	0	2
15.	Iris	150	No	4	0	3
16.	Labor	57	Yes	8	8	2
17.	Lympho	148	No	3	15	4
18.	Mushroom	8,124	Yes	0	22	2
19.	Primarytumor	339	Yes	0	17	21
20.	Sonar	208	No	60	0	2
21.	Vehicle	846	No	18	0	4
22.	Vowel	990	No	10	3	11
23.	Waveform	5,000	No	41	0	3
24.	Zoo	101	No	1	16	7

Any of the interested readers can obtain the details of all the datasets from the UCI Machine Learning Repository [14]. The experimental methodology of 10-fold cross-validation is used for accuracy and tree size results. In 10 fold cross validation, the datasets is split into 10 folds and 9 folds are used for training and 10th fold is used for testing; the same process is repeated by changing the testing fold for 10 runs.

The implementation is done on the open platform Weka [15] on windows 7 with i5 CPU running 3.25 GHz unit with 4 G RAM. A good number of data sources are used for validation of the new proposal against the compared algorithms. All the algorithms are compared with the proposed approach on equal terms. The parameters used for compared algorithms are show in the Table 2.

Table 2. Experimental Settings for Standard Decision Tree Algorithms



Algorithm	Parameter	Value
C4.5	confidence factor	0.25
	min number of objects	2.0
REP	maximum depth	no restriction
	min number of objects	2.0
	min variance proportion	0.001
CART	number of folds pruning	5
	min number of objects	2.0
NB Tree	technique used at leaves	naive bayes

5. Results and Discussions

The experimental validation of the proposed approach was done on 24 datasets from UCI repository. The proposed RMPOD-sparse and RMPOD-Gaussian methods are compared with four classical and well-known decision tree algorithms: C4.5, REP, CART and a well-established NB Tree algorithm. The results of our proposed RMPOD-sparse and RMPOD-Gaussian algorithms are far better than the compared traditional decision tree algorithm. The validation metrics accuracy and tree size are used. Table 3 and 4 presents the experimental results of accuracy ad tree size respectively for both proposed and compared algorithms. The comparative study of the proposed approach is done on independently for every algorithm. This approach is the most used approach for efficient validation of the proposed algorithm. In Table 3, the best two accuracy values in each row are bold faced and one can observe that in most of the cases our proposed RMPOD-sparse and RMPOD-Gaussian algorithms have produced the best accuracy values. In Table 4, the best two tree size values are bold faced and our proposed approaches RMPOD-sparse and RMPOD-Gaussian have performed competitively.

Table 3. Summary of Tenfold Cross Validation Performance for Accuracy on all the Datasets

Datasets	C4.5	REP	CART	NB Tree	RMPOD–sparse	RMPOD-Gaussian
Anneal.ORIG	92.35	91.89	93.36	**97.13**	**98.47**	91.54
Balance-scale	77.82	78.54	78.73	75.96	**95.79**	**96.67**
Breast-cancer	74.28	69.35	70.22	70.99	**91.53**	**98.86**
Breast-cancer-w	95.01	94.77	94.74	96.37	**99.07**	**100.00**
Horse-colic	85.16	84.94	85.37	81.11	**96.14**	**99.4**
Credit-rating	85.57	84.75	84.99	85.42	**96.59**	**96.63**
German_credit	71.25	72.02	73.43	74.64	**94.92**	**99.86**
Pima_diabetes	74.49	74.46	74.56	74.96	**98.52**	**95.81**
Glass	67.63	65.54	71.26	69.84	**80.06**	**80.76**
Heart _c	76.94	77.02	78.68	80.03	**98.04**	**97.52**
Heart-h	80.22	78.56	79.02	81.50	**98.54**	**97.92**
Heart-statlog	78.15	76.15	78.07	80.93	**94.03**	**97.49**
Hepatitis	79.22	78.62	77.10	81.30	**84.45**	**100.00**
Ionosphere	89.74	89.46	88.87	**90.03**	**96.94**	85.62
Iris	**94.73**	93.87	94.20	93.47	**100.00**	93.95
Labor	78.60	78.27	80.03	**91.63**	**90.13**	73.87
Lymphography	75.84	75.33	77.21	81.90	**85.32**	**96.79**
Mushroom	**100.00**	99.98	99.95	**100.00**	98.59	97.34
Primary-tumor	41.39	38.71	41.42	47.50	**78.75**	**63.01**
Sonar	73.61	72.69	70.72	**77.11**	75.21	**86.93**
Vehicle	72.28	70.18	69.91	70.98	**82.56**	**80.07**
Vowel	80.20	66.67	79.61	**92.35**	**83.19**	78.42
Waveform	75.25	76.57	76.65	79.84	**91.05**	**91.11**

Zoo	**92.61**	40.61	40.61	**94.73**	90.59	82.32

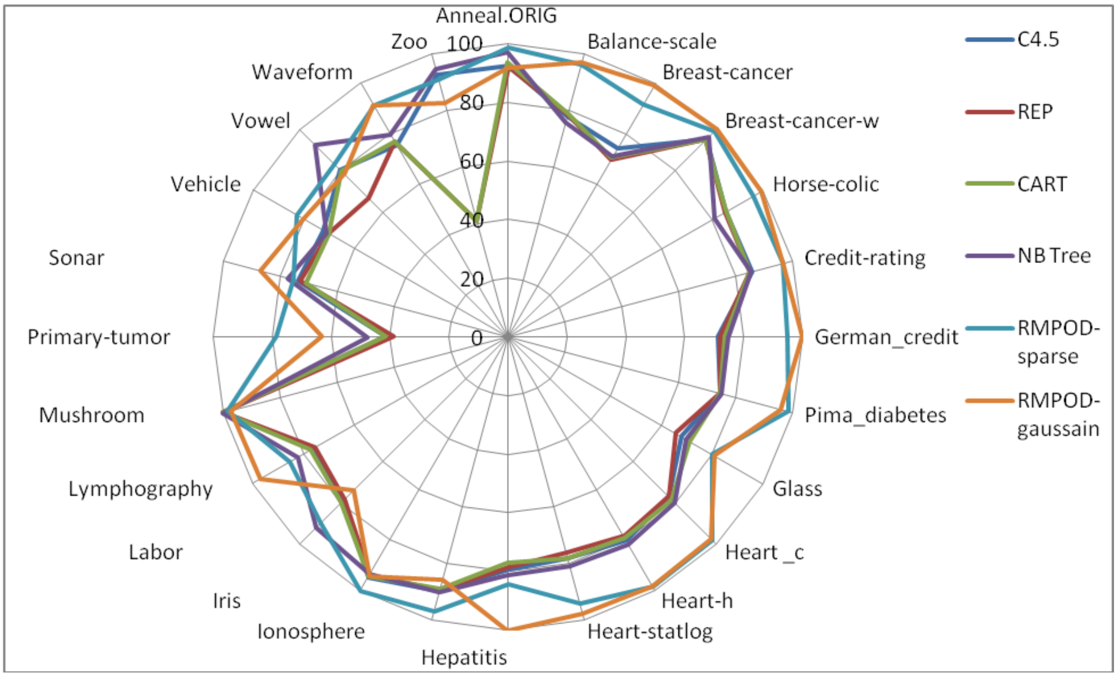


Table 4. Summary of Tenfold Cross Validation Performance for Tree Size on all the Datasets

Datasets	C4.5	REP	CART	NB Tree	RMPOD–sparse	RMPOD-Gaussian
Anneal.ORIG	68.64	63.53	93.22	**32.93**	**49.86**	101.56
Balance-scale	82.20	42.36	55.28	17.38	**55.12**	**22.62**
Breast-cancer	12.78	30.70	7.16	11.90	**26.32**	**3.00**
Breast-cancer-w	23.46	13.76	15.90	**5.68**	12.72	**8.62**
Horse-colic	8.80	15.19	**6.42**	24.27	16.70	**5.00**
Credit-rating	32.82	22.03	**6.54**	17.90	28.32	**13.32**
German_credit	126.85	76.81	24.46	**12.07**	41.12	**5.00**
Pima_diabetes	43.40	30.98	17.36	**5.18**	**13.08**	26.30
Glass	46.16	19.70	**21.16**	**10.0**	31.58	32.94
Heart-c	42.52	18.39	**13.82**	14.58	14.94	**6.58**
Heart-h	**10.53**	13.63	13.42	10.61	14.86	**7.02**
Heart-statlog	34.64	14.78	15.36	**9.62**	25.58	**6.94**
Hepatitis	17.66	**5.64**	6.04	11.56	19.96	**1.00**
Ionosphere	26.74	**8.76**	**8.42**	16.20	18.28	28.04
Iris	8.28	5.84	7.40	**4.38**	**5.00**	8.72
Labor	6.92	**6.15**	9.32	**4.46**	6.96	10.52
Lymphography	28.00	11.46	13.92	**10.24**	18.26	**6.60**
Mushroom	29.94	37.54	**13.24**	**27.55**	219.70	213.24
Primary-tumor	81.51	**33.50**	29.04	**8.79**	47.82	73.42
Sonar	27.90	**10.20**	**10.50**	13.74	26.54	25.82
Vehicle	138.0	**58.52**	92.5	**57.70**	118.20	122.02
Vowel	209.81	254.36	**171.74**	**70.10**	203.28	215.54
Waveform	591.94	167.24	**98.32**	**94.48**	309.98	313.06
Zoo	15.70	**1.00**	**1.00**	8.34	13.72	19.66

The reasons for improved performance of RMPOD-sparse and RMPOD-Gaussian algorithms are due to removal noisy and irrelevant attributes from the data source. The other reason is due to procedural/intellectual learning approach and the final reason is due decrease in the noisy and outliers instances.

The observations from Table 3, 4 and Figure1suggest that:

- Our proposals are the best performing algorithms when the datasets are complex and high dimension.
- The applicability of our algorithm on real time datasets is demonstrated.

The improved result achieved by RMPOD-sparse and RMPOD-Gaussian are due to the efficient orthogonal transform of the feature set. The dimensionality and complexity of dataset is dramatically reduced by the proposed approach and thereby helping the base algorithm C4.5 to perform better on the data source.

Finally, one can observes that the proposed approaches are best suitable when dealing with datasets of complex and high dimensional in nature. The research findings in this study conclude that the orthogonal attribute transform approach can produce better results when dealing with real world datasets.

6. Conclusion

This paper presents a new decision tree algorithm Radom Matrix Projection with Outlier Detection (RMPOD). The proposed algorithms improve the accuracy on high dimensional and complex datasets. In future, the proposed approaches will be extended with other statistical measures.

Acknowledgments

The authors thank all the reviewers for their unanimous comments which improved the quality of the paper.

References

[1] E. Bingham and H. Mannila, “Random projection in dimensionality reduction: applications to image and text data”, KDD'01 Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, **(2001)**, pp. 245-250.

[2] A. S. Jeyalatha and R. Sumbaly, “Diagnosis of diaetiesusig classification miigtechiques”, International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. 5, no. 1, **(2015)**.

[3] M. P. Bach, J. Zoroja and V. Šimičević, “Attribute Selection for Predicting Credit Default with Decision Trees”, Economics and Management Engineering, World Academy of Science, Engineering and Technology, vol. 2, no. 6, **(2015)**.

[4] P. Mondal and K. Mali, “Cardiac Arrhythmias Classification using Decision Tree”, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 1, **(2015)**.

[5] S. Manohar, A. Mittal, S. Naik and A. Ambre, “A Dynamic Classifier using Decision Tree Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 1, **(2015)**.

[6] M. Arif, K. A. Alam and M. Hussain, “Crime Mining: A Comprehensive Survey”, International Journal of u- and e- Service, Science and Technology, http://dx.doi.org/10.14257/ijunesst.2015.8.2.34, vol. 8, no. 2, **(2015)**, pp. 357-364.

[7] K. Fawagreha, M. M. Gabera and E. Elyana, “An Outlier Detection-based Tree Selection Approach to Extreme Pruning of Random Forests”.

[8] N. A. Abhinaya, “An Effective Progress for Extreme Deviation Spying with Flawed Data Labels”, International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 2, **(2015)**.

[9] P. Ajitha1 and Dr. E. Chandra2, “A Survey on Outliers Detection in Distributed Data Mining for Big Data”, Journal Basic Application Science Res., vol. 5, no. 2, **(2015)**, pp. 31-38.

[10] Ms. N. Momin and Prof. N. Hambir, “A Survey on various classification and novel class detection approaches for feature evolving data stream”, Multidisciplinary Journal of Research in Engineering and Technology, vol. 2, no. 1, pp. 342-346.

[11] A. K. Menon, “Random projections and applications to dimensionality reduction”, Honors Thesis, University of Sydney, **(2007)**.

[12] D. Achlioptas, “Database-friendly random projections”, In SIGMOD, **(2001)**, pp. 274–281.

[13] P. Li, T. J. Hastie and K. W. Church, “Very sparse random projections”, In KDD, **(2006)**.

[14] C. Blake and C. J. Merz, “UCI repository of machine learning databases”, Machine-readable data repository. Department of Information and Computer Science, University of California at Irvine, Irvine. http://www.ics.uci.edu/mlearn/MLRepository.html, **(2000)**.

[15] I. H. Witten and E. Frank, “Data Mining: Practical machine learning tools and techniques”, 2nd edition Morgan Kaufmann, San Francisco, **(2005)**.