computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

INTEGRATED EFFECT OF DATA CLEANING AND SAMPLING ON DECISION TREE LEARNING OF LARGE DATA SETS

Dipak V. Patil [1)], Rajankumar S. Bichkar [2)]

[1)] Department of Computer Engineering
Sandip Institute of Technology and Research Centre, Nasik, M.S., India.
dvspatil@yahoo.co.in
[2)] Department of Computer Engineering
G.H. Raisoni College of Engineering & Management, Pune, M.S., India.
bichkar@yahoo.com

Abstract: *The advances and use of technology in all walks of life results in tremendous growth of data available for data mining. Large amount of knowledge available can be utilized to improve decision-making process. The data contains the noise or outlier data to some extent which hampers the classification performance of classifier built on that training data. The learning process on large data set becomes very slow, as it has to be done serially on available large datasets. It has been proved that random data reduction techniques can be used to build optimal decision trees. Thus, we can integrate data cleaning and data sampling techniques to overcome the problems in handling large data sets. In this proposed technique outlier data is first filtered out to get clean data with improved quality and then random sampling technique is applied on this clean data set to get reduced data set. This reduced data set is used to construct optimal decision tree.*

Experiments performed on several data sets proved that the proposed technique builds decision trees with enhanced classification accuracy as compared to classification performance on complete data set. Due to use of classification filter a quality of data is improved and sampling reduces the size of the data set. Thus, the proposed method constructs more accurate and optimal sized decision trees and it also avoids problems like overloading of memory and processor with large data sets. In addition, the time required to build a model on clean data is significantly reduced providing significant speedup.

Keywords: *Large data sets, decision tree, data cleaning, data sampling, speedup and classification accuracy.*

1. INTRODUCTION

The volume of data in databases is growing to large sizes, both in the number of attributes and instances. Data mining provides tools to inference knowledge from databases. This knowledge is used to boost the businesses. Data mining on a very large data set may overload a computer systems memory and processor making the learning process very slow. Data sets used for inference may be very large, may be up to terabytes. The data mining methods are faster when used on smaller data sets. T. Oates, D. Jensen [1] proved that removing randomly selected training instances;often results in smaller trees that are equally accurate to those built on all available training instances. Random data reduction technique is one of the solutions to handle the large data sets. [2].

The noise or outliers in data also affect the classification performance of the classifier on that data. An outlier is an instance that is significantly divergent with rest of the data set. [3]. Gamberger and Lavarac [4] suggested that effect of erroneous data on hypothesis could be avoided by eliminating it from training data before induction.

Data cleaning and sampling reduces time complexity of decision tree learning. Let n be the number of training instances and m be the number of attributes in a instance, computational cost of building tree is $O(m\ n\ log\ n)$ [5]; as n is reduced due to sampling and filterring, the corresponding cost is reduced. Similarly the cost of decision tree pruning process is $O(n(logn)^2)$ [5]. Due to data cleaning operation overfitting can be reduced and as erroneous data is removed the time complexity of pruning process can be reduced significantly.

The proposed technique integrates data cleaning and data sampling techniques to overcome these problems. The classification filter is used to filter training data to improve data quality, subsequently incremental random sampling is applied on this filtered data. These cleaned and sampled data sets

are used to learn classifiers.

1.1 Decision Tree Construction

Decision tree is a classifier in the form of tree that contains a decision node and leaves. A decision node specifies a test to be carried on single attributes value. A leave specifies a class. A solution is at hand for each probable outcome of the test in the form of child node. A tree is traversed from root to node to find out the class of an instance. A performance measure of a decision tree is the number of correctly classified instances called classification accuracy of the tree. It is defined in terms of the percentage of correctly classified instances. [6] - [8]. A decision tree algorithms construct accurate decision trees for classification, but they often experience the drawback of excessive complexity that can make them incomprehensible to human experts [9].

Hybrid learning methodologies that integrate genetic algorithms (GAs) and decision tree learning for evolving optimal decision trees have been proposed by different authors. Although the approaches are different the objective is to obtain optimal decision trees. The decision tree is called optimal if it is accurate and has minimum number of leafs. The GAIT algorithm proposed by Z. Fu [10] proposed generation of the set of diverse decision trees from different subsets of the original data set by using a decision tree algorithm C4.5, on small samples of the data. These decision trees are used as the initial populations to genetic algorithm. The fitness criterion for evaluation is the classification accuracy on test data. A. Papagelis, and D. Kalles proposed GATree, a genetically evolved decision tree [11]. The Genetic Algorithm is used to directly evolve binary decision trees. Decision trees those have one decision node with to two different leaves are operated by genetic operators. The constructed trees are called genetically evolved decision trees. Similar approaches are available in [12] - [13].

2. RELATED WORK

The proposed approach is combination of two techniques first is data cleaning and second is data sampling, these techniques are presented in two separate subsections.

2.1 DATA CLEANING

The outlier detection and noise elimination is an important issue in data analysis. The exclusion of outliers improves data quality and therefore classification performance. Several researchers have proposed various approaches for data cleaning. G.H. John [14] proposed a technique that removes a misclassified training instance from training data and reconstructs the trees, the process is repeated till all such instances are removed from training data. Misclassified instances are recognized using tree classifier as a filter. The resulting classifier enhances classification performance accuracy. Broadly and Friedl [15] proposed a method for detecting mislabeled instances. The method uses a set of learning algorithms to construct classifiers that act as a filter for the training data. The technique removes outliers in regression analysis. Arning et al. [16] proposed framework for the problem of outlier detection. Similar to human beings, it observes all instances for similarity with data sets and it treats dissimilar data set as an exception. A dissimilarity function is used to find out outliers.

Raman and Hellerstein [17] proposed an interactive framework for data cleaning that integrates transformation and discrepancy detection. Guyon et al [18] proposed training of convolutional neural networks with local connections and shared weights. Gamberger and Lavrac [19] proposed conditions for Occam's razor applicability in noise elimination. Knorr and Ng [20] proposed unified outlier detection system. Subsequently, they proposed and analyzed some algorithms for detecting distance-based outliers. Tax and Duin [21] proposed outlier detection that is based on the instability of the output of simple classifiers on new objects. Gamberger et al [22] proposed saturation filter. It is based on principle that detection and removal of noisy instances from training data induces less complex and more accurate hypothesis. The saturated training data set can be employed for induction of stable target theory.

Schwarm and Wolfman [23] proposed Bayesian methods for data cleaning which detects errors and corrects them. Ramaswamy et al [24] proposed algorithm for distance-based that ranks each point based on its distance to its nearest neighbor. Kubika and Moore [25] presented system for learning explicit noise. The system detects corrupted fields and uses non-corrupted fields for consequent modeling and analysis. Verbaeten and Assche [26] proposed three ensemble based methods for noise elimination in classification problems. Loureiro et al [27] proposed a method that applies hierarchical clustering methods for outlier detection. Xiong et al. [28] proposed a hyperclique-based noise removal system to provide superior quality association patterns. Patil and Bichkar [29] proposed use of evolutionary decision tree as classification filter and found that, with use of genetic algorithm optimal trees can be built.

2.2 DATA SAMPLING

Sampling is the procedure to obtain a subset of

instances that represents the entire data set. It is necessary to have sufficient sample size to validate statistical analysis. Sampling is done because it is impracticable to test every single individual in the data set. Moreover it saves time, resources and effort while executing the research. The representativeness is the most important issue in statistical sampling. The sample obtained from the set of data instances must be representative of the same. Probability sampling and Non-probability sampling are two types of sampling. In Probability sampling, all the instances in set have equal probability of being selected. The approach assures completely randomized selection process which is unbiased. The hypothesis is accurate when this sampling is used. In Non-Probability sampling all the instances in data set do not have equal probability of being selected. Thus sample does not completely represent the target data set. The representativeness can be achieved by using simple randomized statistical sampling techniques.

Jenson and Oates [1] experimented with random data sampling and proved that as size of the training dataset increases size of tree also increases where as classification accuracy does not increase significantly.

S. Vucetic and Z. Obradovic [30] proposed an effective data reduction method founded on guided sampling for determining a minimal size representative subset, it was followed by a model-sensitivity analysis for determining a suitable compression level for each attribute.

A. Lazarevic and Z. Obradovic [31] proposed several efficient techniques based on the idea of progressive sampling. The sampling process combines all the models constructed on previously considered data samples. The authors also proposed controllable sampling based on the boosting algorithm, where the models are combined using a weighted voting. Another contribution by authors is sampling procedure for spatial data domains, where the data instances are selected in accordance with the performance of previous models as well as in accordance with the spatial correlation of data.

Patil and Bichkar [32] proposed use of evolutionary decision tree along with random sampling of data to optimize the decision trees and concluded that the proposed method builds trees that are accurate and relatively smaller in size.

3. PROPOSED MODEL

The proposed model is based on combination of removal of outliers from data as first stage and incremental random sampling of data to evolve decision trees as second stage to obtain compact representation of large data set. The data set is first filtered using classification filter and then sampled randomly in different subsets by random sampling, initially, sample size is 5% and reaches up to 100% of the instances available in clean data set. The size of training data set is increased by 5% each time using random selection of instances and using this data tree are evolved. We have set 5% sample size as minimum Threshold size. The tree is first on built sample of size 5% of clean data and the classification accuracy on clean sampled data is compared with accuracy on complete clean data set. If classification accuracy on sample data is equal or greater than accuracy complete data then, the sample size is representative of complete data set, otherwise next incremental sample size is 10% of data and so on.

Let T_F be a set of all available n training instances classification filter is applied and we get clean data set T_C and unclean data set T_E. Let X_C is classification accuracy on T_C. Let a training instance be denoted by I and let the sampling size Threshold be denoted by α, and $\alpha = 5$. Random sampling is done on T_C, let T_i be subset of instances in T_C where $T_i \in T_C$ and $i = 1$ to 20, let hypothesis H_i is induced on T_i and classification accuracy on this training data be X_i.

$\forall T_i$ on H_i if $X_i \geq X_C \Rightarrow T_i$ is Final training set for constructing classifier and is denoted as reduced data set T_R and accuracy on T_R is denoted by X_R.

The proposed algorithm works as follows**.**

1. Induce*(H, T_F)*.
2. Classify *(H, T_F)*. // Classification Filter
3. $\forall I$ in T_F If $X(I) = 1 \Rightarrow I \in T_C$.
4. Else del*(I)*.
5. Sample*(T_C, T_i)* //Random sampling on T_C.
6. For *($i = 1$; $i \leq 20$; $i = i+1$)*
7. Induce*(H_i, T_i)*.
8. Compare*(X_i, X_C)*.
9. If $X_C \geq X_i \Rightarrow T_i$ is final reduced data set T_R.
10. Induce hypothesis H_R on T_R with decision tree as a final classifier.
11. Else repeat step 7 to 10 with next incremental sample until we get $X_C \geq X_i$.
12. End.

4. METHOD OF EXPERIMENTATION

Prior to applying the proposed technique on large data sets, we found it appropriate to first test it on the normal sized benchmark data set from UCI repository [33]. Experiments were performed on 24 benchmark data sets to explore classification accuracy of decision tree at reduced training data using proposed approach and results are validated using standard decision tree algorithm J48, CART [34] and GATree[11]. A significant enhancement in

classification performance on all data sets is observed, in order to make the presentation concise we are presenting only few cases with lower, moderate and higher classification accuracy on sample basis; these results are presented in Table 1, 2 and 3 respectively. For these data sets value of α = 25; i.e. initial sample size is 25% as data sets are smaller to midsize data sets. Whenever average results are presented in result discussion, it means it is average of all 24 data sets.

In test method, the data set is first filtered using classification filter. Then clean data is incrementally sampled in different subsets and trained using decision trees algorithms until we get classification accuracies comparable to that obtained on complete data set. The same method without filtering data is also followed in experimentation to get results with unclean data.

In order to analyse effect of cleaning on data and effect of sampling on data we have separate subsections followed by analysis on combined effect of cleaning and sampling on data by proposed method.

4.1 EFFECT OF CLEANING ON DATA

Data cleaning process improves the quality of data. This section presents effect of data cleaning on classification performance of the classifier on clean data sets. To study the effect, the complete data set T_F is filtered using classification filter and clean data T_C is obtained. The classification accuracies on data sets T_F and T_C are obtained using k-fold cross validation method and are presented in Table 1; X_F is classification accuracy on T_F. It is observed that for J48 classifier, classification accuracies on cleaned data sets in Table 1 are around 99% except Monks data set. Monks data set is having lower classification accuracy on complete data set T_F amongst all, it is 70.16% and classification accuracy on cleaned data is 90.16% which is a significant enhancement in classification performance on cleaning data. Similar results are observed on CART classifier. In case of GATree classifier, exceptional case is data set Mfeat-Factor, it has lower classification accuracy on complete data set which is 38.80% and is significantly enhanced to 80.67% with 41.87% enhancement in accuracy. The enhancement in accuracy ΔX is the absolute difference in accuracy between the tree build from the cleaned data set and the tree build from complete data training data which is unclean data. The average results for enhancement in accuracy (average ΔX) due to data cleaning for all 24 data sets are 8.87%, 8.43%, and 25.49% for J48, CART and GATree respectively. The previous results available so far by G. H. John [12] indicate enhancement in accuracy on J48 by 2% to 4% in average whereas here we have enhancement of around 8% in accuracy. Thus, with data cleaning with proposed we get enhanced classification performance, the reason is removal of anomalies in data.

Table 1. Effect of cleaning on data

Sr. No.	Data Set	J48			CART			GATree		
		X_F	X_C	ΔX	X_F	X_C	ΔX	X_F	X_C	ΔX
1	Australian	90.72	100.00	9.28	91.01	100.00	8.99	87.971	100.00	12.02
2	Breast –w	94.85	98.21	3.36	96.42	98.51	2.09	94.10	99.40	5.30
3	German	79.80	99.72	19.92	85.50	99.44	13.94	70.20	99.58	29.38
4	kr-vs-kp	99.41	100.00	0.59	99.62	100.00	0.38	91.05	98.55	7.50
5	Mfeat-Factor	93.50	99.73	6.23	93.25	98.66	5.41	38.80	80.67	41.87
6	Monks	70.16	90.16	20.00	64.52	93.44	28.92	70.00	90.00	20.00

4.2 EFFECT OF SAMPLING ON DATA

As it is impracticable to test every single individual instance in the data set, tests are conducted on samples of data. In this section we present analysis effect of sampling on data, and hence data cleaning process is not exercised here. The results are presented in Table 2. The complete data set T_F is sampled. Initial sample size is 25% of data, train decision tree on it, if classification accuracy on this classifier is equal or more than classification accuracy on complete dataset; the sample size is final sample size T_{RU}. Otherwise increment sample size by 5% and repeat the process. The data is incrementally sampled in different subsets and trained using decision trees algorithms successively until we get classification accuracies comparable to that obtained on complete data set. The sample size is abbreviated as SS and classification accuracy on reduced unclean data set T_{RU} is abbreviated as X_{RU} in Table 2. Here the enhancement in accuracy ΔX is the absolute difference in accuracy between the tree build from the reduced unclean data set and the tree build from complete data training data which is unclean data.

The average classification accuracies on all 24 complete datasets T_F are 90.02%, 89.31% and 72.25% as compared to accuracy of 90.32%, 90.05% and 74.75% on sampled data set T_{RU} for J48, CART

and GATree classifiers respectively. Thus, with data sampling, we get comparable accuracies on reduced data set.

The average percentage sample sizes for all 24 data set are 72.70%, 74.17% and 33.33% on complete data sets for J48, CART and GATree classifiers respectively. The sample size on J48 and CART are similar in average, whereas it is significantly smaller for GATree. Although average sample size is between 72% to 74% for J48 and CART, sampling was not so successful on some unclean data sets. The exceptional cases are, Monks data set on J48 classifier, German and Mfeat-Factor data sets on CART. These data sets could get required comparable accuracy with 100% data samples on unclean data set. The anomalies present in the data set are the most probable reason for it.

The sample size required for GATree is smaller because, GAtree incorporates genetic algorithm for global search in the problem space with classification performance in terms of accuracy as fitness function without being biased towards a local optimum and the sample size required is optimised. Thus we get reduced data set with comparable classification performance.

Table 2. Effect of sampling on data

Sr. No.	Data Set	J48				CART				GATree			
		X_F	X_{RU}	ΔX	SS	X_F	X_{RU}	ΔX	SS	X_F	X_{RU}	ΔX	SS
1	Australian	90.72	90.72	0.00	85	91.01	91.52	0.51	65	87.97	87.50	-0.47	35
2	Breast –w	94.85	95.22	0.37	45	96.42	97.42	1.00	50	94.10	97.65	3.55	25
3	German	79.80	80.15	0.35	65	85.50	85.50	0.00	100	70.20	74.40	4.20	25
4	kr-vs-kp	99.41	99.30	-0.11	40	99.62	99.43	-0.19	55	91.05	91.07	0.02	25
5	Mfeat-Fact.	93.50	93.00	-0.50	75	93.25	93.25	0.00	100	38.80	41.00	2.20	25
6	Monks	70.16	70.16	0.00	100	64.52	67.50	2.98	65	70.00	70.00	0.00	90

4.3 EFFECT OF CLEANING AND SAMPLING ON DATA

In the experimentation with the proposed approach, the data is first filtered and then sampled incrementally and trained using decision trees algorithms until we get classification accuracies comparable to that obtained on clean data set T_C. The comparison of accuracies on complete data set T_F with cleaned and reduced data set T_{RC} is presented here in Table 3. The enhancement in accuracy ΔX is the absolute difference in accuracy between the tree build from the reduced clean data set and the tree build from complete data training data which is unclean data.

The accuracies on clean and reduced data set T_{RC} are around 99% for all data set in Table 3 for J48 and CART. Exceptional case is Monks data set, where we could get around 93% accuracy with enhancement in accuracy of around 23% and 28% for J48 and CART respectively. Monks data set has lower accuracy on T_F hence, there is scope for enhancement. The data set kr-vs-kp is having around 99% accuracy on T_F and hence there is very less scope for enhancement in classification performance on the classifiers J48 and CART.

Similarly in case of GATree accuracies on reduced data sets are 99% on all data sets in table with exceptional case Mfeat-Factor data, where the minimum accuracy is 38.80% which is enhanced to 83.78%.

The average enhancement in accuracy for all 24 data set is 9.28%, 8.84%, and 22.07% for J48, CART and GATree; we get significant enhancement in accuracy with proposed approach.

In analysis of sample data size or reduced data set size, we found that data size in terms of number of training instances gets reduced and required sample size varies from 25% to 45% for J48, CART and GATree. Exceptional case is Breast-w on J48 with 70% sample size.

The average percentage sample data set size for all 24 clean data sets T_{RC} are 38.13%, 39.38% and 30.42% as compared to 72.70%, 74.17% and 33.33% for for unclean data set T_F on J48, CART and GATree respectively. We could observe very less deviation in sample size for GATree classifier from T_F to T_{RC}, the reason is already mentioned in above section.

The numbers of training instances required are less for clean data set as compared to unclean data set. As anomalies are removed the numbers of instances required for induction are also less. Thus with proposed approach we get reduced training data set for induction and upon induction enhanced classification performance is available.

Table 3. Effect of cleaning and sampling on data

Sr. No.	Data Set	J48				CART				GATree			
		X_F	X_{RC}	ΔX	SS	X_F	X_{RC}	ΔX	SS	X_F	X_{RC}	ΔX	SS
1	Australian	90.72	100.00	9.28	25	91.01	100.00	8.99	25	87.97	100.00	12.03	25
2	Breast –w	94.85	99.15	4.3	70	96.42	98.01	1.59	45	94.10	100.00	5.9	30
3	German	79.80	100.00	20.2	25	85.50	100.00	14.5	25	70.20	99.64	29.44	40
4	kr-vs-kp	99.41	100.00	0.59	35	99.62	100.00	0.38	40	91.05	98.67	7.62	25
5	Mfeat-Fact	93.50	99.66	6.16	40	93.25	98.93	5.68	25	38.80	83.78	44.98	25
6	Monks	70.16	93.33	23.17	25	64.52	93.33	28.81	50	70.00	100.00	30	25

5. EXPERIMENTS WITH LARGE DATA SET

The validation of proposed method on big data was done with test on two big benchmark data sets namely, Census-income (KDD) and KDDCup99 data set with data set size of 299,000 and 494,020 instances of data respectively. The classifiers used are J48, CART and GATree. The method of experimentation on big data has exception that the data sampling threshold is 5% instead of 25%. The threshold value for minimum sampling size is 5% for big data sets because the adequate data samples are available for training even at lower sampling percentage with big data set.

5.1. ANALYSIS ON ACCURACY OF THE TREE AND SIZE OF TRAINING DATA

Table 4 presents a comparison of percentage sample size, accuracies on clean and unclean data sets with and without sampling. It also presents the enhancement in accuracy ΔX. Here ΔX is the absolute difference in accuracy between the tree build from the cleaned data set and the tree build from complete data training data which is unclean data. The classification accuracies obtained on clean data T_C are higher, and are around 99% on clean data set, similar results are obtained clean sampled data T_{RC} and thus, classification performance of the proposed method in terms of accuracy is enhanced.

It its observed that the average percentage sample size is lower on clean data set as compared to unclean data. Due to filtering of data, outlier data is removed from data set and as anomalies in data are removed and the data set size required to build the model is also reduced. In case of unclean data the average sample set size around 24% where as for clean data set it around 13%. Thus, significant reduction in data set size is achieved using proposed technique.

Table 4. Comparison on percentage sample size and enhancement in accuracy

Data Set	Classifier	SS		X_F	X_{RU}	X_C	X_{RC}	ΔX
		Unclean	Clean					
Census-income	J48	30	10	95.39	95.26	100.00	99.99	4.51
	CART	30	10	95.50	95.24	100.00	99.99	4.49
	GATree	25	15	93.70	94.23	99.99	99.99	6.29
KDDCup99	J48	25	10	99.96	99.91	99.99	99.98	0.02
	CART	25	10	99.95	99.90	99.99	99.99	0.04
	GATree	10	25	88.56	93.67	98.06	98.30	9.74
Average		**24.17**	**13.33**	**95.51**	**96.37**	**99.67**	**99.71**	**4.18**

Table 5 provides comparison of number of instances in T_F and reduced data set T_{RU}, after cleaning of number of instances in T_C, and reduced data set T_{RC}.

In this table N_F indicates number of training instances in T_F, similarly N_C, N_{RU} and N_{RC} for T_C, T_{RU} and T_{RC}. The Table shows percentage reduction in required training set size ΔN with proposed method. It is the difference in number of training instances in unclean complete data set and cleaned, reduced data set. Where

$$\Delta N = (N_F - N_{RC}) * 100 / N_F \quad (1)$$

As anomalies are removed the required sample size is also reduced for clean data and the reduction is around 90%. This is the most significant achievement with proposed method because the size of data set is reduced considerably along with enhanced classification performance; the reduced data set help to deal with the memory and processor limitations. Thus reduction in training set size achieved is significant.

Table 5. Comparison data reductions and speed up in case unclean and clean large data sets.

Data Set	Classifier	Unclean Data				Clean Data				Δ t	Δ N
		N_F	$t(T_F)$	N_{RU}	$t(T_{RU})$	N_C	$t(T_C)$	N_{RC}	$t(T_{RC})$		
Census	J48	299K	137.84	89.7K	17.78	224.1K	19.88	22.4K	2.98	97.84	92.51
	CART	299K	8521.48	89.7K	1440.48	224.1K	1470.23	22.4K	79.24	99.07	92.51
	GATree	299K	1593.00	74.7K	352.00	224.1K	252.00	33.6K	32.00	97.99	88.76
KDDCup	J48	494K	305.70	123.5K	30.67	280.0K	48.17	28.0K	2.20	99.28	94.33
	CART	494K	1791.00	123.5K	242.22	280.0K	368.2	28.0K	23.84	98.67	94.33
	GATree	494K	2523.00	49.4K	550.00	280.0K	492.00	70.0K	60.00	97.62	85.83

5.2. ANALYSIS ON TIME REQUIRED TO BUILD THE MODEL

In this experimentation with large data sets one more parameter «time required to build the model» is added for analysis as it is significant in case of large data sets. In this analysis $t(T_F)$ indicate time required to build model on T_F, similarly $t(T_C)$, $t(T_{RU})$ and $t(T_{RC})$ indicate time required to build model on T_C, T_{RU} and T_{RC}.

The speedup or percentage reduction in average time required to build the model Δt with proposed method is the difference between average time required to build the model with all available data instances T_F and average time required to build the model on cleaned and reduced data set T_{RC}. The percentage of reduction in time required to build the model is given by

$$\Delta t = (t(t_F) - t(t_{RC})) * 100 / t(T_F) \qquad (2)$$

The results are presented in Table 5. Due to filtering of data, as anomalies in data are removed, time complexity of tree pruning process $O(n(\log n)^2)$ is reduced and thus decision tree learning process is accelerated. Here it is observed that the time required to build the model on clean data set is significantly lower than the time required building model on unclean data of same size. When we combine data cleaning and data sampling, time complexity of tree induction process $O(m\ n \log n)$ is also reduced due to reduction in n, significant reduction in time to build the model is observed for reduced clean data set T_{RC} as compared to complete data set T_F and there is around 98% reduction in time required to build the model for all cases in Table 5. Thus significant speedup is achieved with the proposed method.

6. CONCLUSION

The proposed approach is a combination of removal of noise from training data and incremental random sampling. The aim is to evolve decision trees in order to obtain compact representation of large data sets. The removal of outlier data improves the quality of training data. The results with this combined approach indicate significant improvements in classification performance along with data reduction.

The possible reduction in sample size depends on nature of data set and it cannot be fixed, but with significant improvement in data quality, significant reduction in required sample size with reference to complete data set is observed. The problems like memory and processor overloading can be handled due to reduced data set.

Since, noise is removed from training data, the time required to build the model is also reduced. Overfitting occurs due to noise in data. For n training instances with m attributes computational cost of building tree is $O(m\ n \log n)$, as n is reduced due to sampling and filtering the corresponding cost is reduced. Similarly the cost of decision tree pruning process is $O(n(\log n)^2)$. Due to data cleaning operation overfitting can be reduced and the time complexity of pruning process is reduced significantly. It is verified that the time required to build model on clean data is very low as compared to time required to build the model on unclean data of same size. The proposed approach improves data quality and reduces data set size while maintaining the classification performance and by virtue of improved data quality and reduced data size the proposed method gives excellent speedup. In conclusion with proposed method the size of data set is reduced, classification accuracy is improved, speedup is acquired and the problem of limitations of memory and processor can also be solved.

7. REFERENCES

[1] Tim Oates, David Jensen, The effect of training set size on decision tree complexity, *Proceedings 14th International Conference on Machine Learning*, (1997), pp. 254-262.

[2] G.H. John and Pat Langley, Static versus dynamic sampling for data mining, *In Proceedings of the Second International*

Conference on Knowledge Discovery in Databases and Data Mining, 1996.

[3] V. Barnett and T. Lewis, *Outliers in Statistical Data*, John Wiley and Sons, 1978.

[4] Gamberger, N. Lavrac, and S. Dzeroski, Noise elimination in inductive concept learning: a case study in medical diagnosis, *In proceedings of 7th International Workshop on Algorithmic Learning Theory*, Sydney, 1996.

[5] Ian H. Witten and Eibe Frank, *Data Mining Practical Machine Learning Tools and Techniques*, Morgan Kaufmann publications, 2005.

[6] Quinlan J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufman, San Mateo, 1993.

[7] Quinlan J.R., Decision trees and decision making, *IEEE Transaction on Systems, Man, and Cybernetics,* (20) 2 (1990), pp. 339-346.

[8] Salvatore Ruggieri. Efficient C4.5: IEEE Transaction On Knowledge and Data Engineering, (14) 2 (2002), pp. 438-444.

[9] Quinlan J.R., Simplifying decision trees, International Journal of Man Machine Studies, 1987.

[10] Zhiwei Fu, Fannie Mae, A computational study of using genetic algorithms to develop intelligent decision trees, *Proceedings of the 2001 IEEE congress on evolutionary Computation*, 2001.

[11] Athanassios Papagelis, Dimitrios Kalles, GATree: genetically evolved decision trees, *Proceedings 12th International Conference on Tools with Artificial Intelligence*, (13-15 November 2000), pp. 203-206.

[12] A. Niimi and E. Tazaki, Genetic programming combined with association rule algorithm for decision tree construction, *Proceedings of Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, 2 (2000), pp. 746-749.

[13] Y. Kornienko and A. Borisov, Investigation of a hybrid algorithm for decision tree generation, *Proceedings of the Second IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing*, 2003.

[14] G. H. John, Robust decision trees: removing outliers from databases, *In Proceedings of the First ICKDDM*, (1995), pp. 174–179.

[15] C. E. Brodley and M. A. Friedl, Identifying mislabelled training data, *Journal of Artificial Intelligence Research*, (1999), pp. 131-167.

[16] Arning, R. Agrawal, and P. Raghavan, A linear method for deviation detection in large databases, *In KDDM*, 1996, pp. 164-169.

[17] V. Raman and J.M. Hellerstein, An interactive framework for data transformation and cleaning, *Technical report University of California Berkeley*, California, September 2000.

[18] A. I. Guyon, N. Matic and V. Vapnik, Discovering informative patterns and data cleaning, *Advances in knowledge discovery and data mining, AAAI*, (1996), pp. 181-203.

[19] D.Gamberger and N. Lavrac, Conditions for Occam's razor applicability and noise elimination, *Proceedings of the 9th European Conference on Machine Learning*, Springer, (1997), pp. 108-123.

[20] E. M. Knorr and R. T. Ng, Algorithms for mining distance-based outliers in large datasets, *In Proceedings 24th VLDB*, (1998), pp. 392-403.

[21] D. Tax and R. Duin, Outlier detection using classifier instability, *Proceedings of the workshop Statistical Pattern Recognition*, Sydney, 1998.

[22] D. D. Gamberger, N. Lavrac, and C. Groselj, Experiments with noise filtering in a medical domain, *In Proceedings 16th ICML*, Morgan Kaufman, San Francisco, CA, (1999), pp. 143-151.

[23] S. Schwarm and S. Wolfman, Cleaning data with Bayesian methods, *Final project report for University of Washington Computer Science and Engineering CSE574*, March 16, 2000.

[24] S. Ramaswamy, R. Rastogi, and K. Shim, Efficient algorithms for mining outliers from large data sets, *ACM SIGMOD*, (29) 2 (2000), pp. 427-438.

[25] J. Kubica and A. Moore, Probabilistic noise identification and data cleaning, *Third IEEE International Conference on Data Mining*, (19-22 Nov. 2003).

[26] V. Verbaeten and A. V. Assche, *Ensemble Methods for Noise Elimination in Classification Problems*, In Multiple Classifier Systems. Springer, 2003.

[27] J. A. Loureiro, L. Torgo, and C. Soares, Outlier detection using clustering methods: a data cleaning application, *In Proceedings of KDNet Symposium on Knowledge-based Systems*, Bonn, Germany, 2004.

[28] H. Xiong, G. Pande, M. Stein and Vipin Kumar, Enhancing data analysis with noise removal, *IEEE Transaction on knowledge and Data Engineering*, (18) 3 (2006), pp. 304-319.

[29] D. V. Patil and R. S. Bichkar, Improving classification performance of evolutionary decision tree using classification filter, *Third International Conference on Info. Processing*, Bangalore, India, (2009), pp.696-700.

[30] Slobodan Vucetic and Zoran Obradovic, Performance controlled data reduction for

knowledge discovery in distributed databases, *Proceedings 4th Pacific-Asia Conference, PADKK 2000*, Kyoto, Japan, (18-20 April 2000), pp. 29-39.

[31] A. Lazarevic and Z. Obradovic, Data reduction using multiple models integration: principles of data mining and knowledge discovery, *5th European Conference, PKDD 2001*, Freiburg, Germany, (3-5 September 2001), pp. 301-313.

[32] D. V. Patil and R. S. Bichkar, A hybrid evolutionary approach to construct optimal decision trees with large data sets, *In Proceedings IEEE ICIT06*, Mumbai, India (15-17 December 2006), pp. 429-433.

[33] Frank A. and Asuncion A., UCI Machine learning repository Irvine. CA, [http://archive.ics.uci.edu/ml]. University of California, School of Information and Computer Science, 2010.

[34] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, *The WEKA Data Mining Software: An Update; SIGKDD Explorations*, (11) 1 (2009).

Dipak V. Patil *received B.E. degree in Computer engineering in 1998 from University of North Maharashtra India and M. Tech. degree in Computer Engineering in 2004 from Dr. B. A. Technological University, Lonere, India. Currently he is an Associate Professor in Computer Engineering Department at Sandip Institute of Technology and Research Centre Nasik, India. His research interests are in data mining specifically handling noise in data and Handling large data sets.*

Rajankumar S. Bichkar *received B.E. (Electronics Engineering) and M.E. (Electronics Engineering) degrees from S.G.G.S. Institute of Engineering and Technology, Nanded India in 1986 and 1991 respectively, and Ph.D. degree in Engineering from prestigious Indian Institute of Technology, Kharagpur India in 2000.*

He has published his research papers in various International Journals. Currently he is a Professor (Electronics and Telecommunication Engineering) and Dean (R&D) at G.H. Raisoni College of Engineering & Management, Pune, India. His research interests are in genetic algorithms, image processing, scheduling, timetabling and data mining.