Journal of Theoretical and Applied Information Technology
30th September 2014. Vol. 67 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OPTIMIZED PRUNE BASED DATA MINING (OPBDM) FOR DISTRIBUTED DATABASES: AN ADAPTIVE APPROACH

G.PARTHASARATHY[1], Dr.D.C.TOMAR[2]

[1]Research Scholar, Dept. of Computer Science and Engineering, Sathyabama University, Chennai, India.

[2]Senior Professor, Dept. of Information Technology, Jerusalem College of Engineering, Chennai, India.

E-mail: [1]amburgps@gmail.com, [2]dctomar@gmail.com

ABSTRACT

Data mining is a vital solution approach for applications such as text based mining; web based mining, supporting data analysis, data visualization and data design tasks. It serves as primary, as well as major roles in various domains and plays a complementary role in others by increasing traditional data mining techniques from numerical data analysis, data statistics and machine learning. Modern mining algorithm concentrates mainly on developing, building and using specific cross platforms, domain driven algorithms and models. In such cases various algorithms are designed in which very few are repeatable in real time. Countless outlines are mined and major proportion of them is of no specific interest for business use. The end user generally cannot easily predict or understand to take them over for business values and daily use. Thorough efforts are mandatory for promoting the act of knowledge discovery in real world decisions and decision supports. An optimized method has been proposed for pruning based data mining to handle the issues stated above. In optimized prune based data mining (OPBDM), Knowledge intelligence is incorporated into mining process along with models and a problem solving system is framed as the gallery for incorporating knowledge discovery, decision making, data pruning and delivery. Based on related work(Prune tree), this paper presents an overview of deriving prunes for multiple data sets, decisions and decision supports, theoretical framework for pruned data(OPBDM)techniques, case studies and open issues. Experimental results and future work enhancement have been demonstrated.
Keywords: *Data Mining, Knowledge Intelligence, Knowledge Discovery, Pruning, Decision Making*

1. INTRODUCTION

Current advancement in computer, computational intelligence, communication, and digital storage technologies, mobile oriented technologies, sophisticated service oriented architecture based technologies (mobile, remote cloud) have seen the procurement of voluminous data to be accessible remotely across the inter and intra domains geographical and administrative boundaries. There is an increasing demand on data mining over the distributed data stores to find the patterns or rules that benefit all of the business participants and clients. For example, multiple retailer stores in the same business section want to cluster the data together (distributed dataset) to determine the characteristics of customer purchases. However, these distributed datasets could also contain sensitive information, such as business sales data, journal citation data, product value sets and patient clinical records in major domains. Therefore, an important challenge for distributed dataset in mining is how to protect each participant's sensitive information, while still finding useful data models (for example classification models). Classification is an important area in data mining and DM based research. Given a set of data, records, each of which outbursts to one of a predefined packages, the classification (Decision tree based) problem fully focused with the discovery of classification and classifiers. Protocols that can allow records, data sets with known and unknown class association are to be properly classified. An optimized method has been introduced for pruning based classification algorithm to handle the multiple data sets, decisions and decision supports.

2. RELATED WORKS

Since many algorithms have been established to mine huge data sets for classification (Decision based) models, the operational features of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

protobuilds have become patent [30], [16], [3], [17], [36]–[38]. Current data mining algorithms such as decision tree based algorithms (e.g., BOAT [16], C4.5 [36], PUBLIC [37], Rain-Forest [17], SLIQ [30], SPRINT [38]) can be used to discover classification rules for classifying records with known classes, unknown classes and class membership. During the course of extension of algorithms using a decision tree for regulation of the possibility relating to such association, the likelihood of similar class likelihood did exist. The algorithms are built on the concepts of rough set theory [44], cluster (data) analysis [1], and measure theory [5]. Computational analysis which indicates the proposed algorithms offer frequent advantages over other approaches such as soft computing based (ANN) and regression analysis [39], namely:

- Simplicity;
- High accuracy;
- Low computational complexity.

An adaptive feature extraction approach is used here which is based on an "individual data object based method" [42]. A feature extraction algorithm recognizes unique structures (business data sets, reputed customers, valued clients, business records etc.) of an object (e.g., business cart) and checks the sharing of the unique features with other objects (spatial database data sets or records and documents). It is obvious that the "population based" and "individual based" patterns differ and, in general, the set of features derived by each of the two patterns is different. In the feature extraction approach, a set of features applies to a group of objects. These features are expressed as a decision rule. The properly derived decision rules (made by feature sets) perfectly assign results for a large proportion of cases with known and unknown decisions (it makes predictions for new cases). The drawback of the feature extraction approach is high complexity of the knowledge phase; however, the computational complexity is reduced through pruning the data sets and records which offers a grander promise for applications in decision-making(Decision tree based) than any of the "population" based approaches.

The method presented in this paper follows the emerging concepts from the rough set theory [12] of data mining. The perceptive behind the rough set theory is that a group of objects (business carts, business data sets, reputed customers, valued clients, business records) with a unique subset of features which shares the same sets of decision outcomes and results. The feature extraction algorithm dynamically analyses a huge database and recognizes unique features of each group of object and evaluates it with reduced inputs for prune tree where the pruning tree decides the process and evaluates it with reduced computational complexity and high objectivity (prediction of reputed customers, product sets). Other classification techniques (Decision tree based) such as regression and Soft computing (ANN) [3] can determine a probability for a prediction with its prospect (likelihood). However, comparing with decision tree based algorithms; these algorithms do not perceptibly direct the exposed patterns in a symbolic (Literals), and easily logical form which is if-then rules [43].

In this paper (OPBDM (Pruning based data mining)); the focus is on award of organised approach and overview of concepts, frameworks, methodologies and techniques of OPBDM. In Section 3 the focus is on pruning data in existing data mining. In Section 4 a multidimensional requirement for knowledge discovery for pruned data through reduction of computational complexity is presented. In Section 5 discusses the concept and basic framework of OPBDM. Section 6 deals with refinement of reduction of error rate in pruning and conclusions are given in section 7.

3. PRUNING DATA

With the vast advancement in digital recording devices and multimedia devices, rapid development of internet platform, and the low cost of storage devices such as micro card, macro card, SD card, Hard disk drives, it becomes much easier to distribute and collect the multimedia data nowadays. The rapid increase and development in the volume of the multimedia data, the inefficient outdated text-based data/information retrieval approaches, the great demands for the multimedia information/data sets analysis and management apps have inspired the researchers to dwell on into the area of "content-based multimedia retrieval from spatial databases"[17][45]. To address the data imbalance issue, the *data pruning [45]* process/technique can be utilized in the manner that given the training data set and the learning model, it can reduce the data set and select the representative data instances as the new training data set so that the performance of the model and the classification result would be improved [45]. For traditional document retrieval, information filtering is either content-based or collaborative one. The content based method is usually based on the term frequency of text documents and the collaborative filtering method is based on the particular user's selection [45]. We use classification method is used

Journal of Theoretical and Applied Information Technology
30th September 2014. Vol. 67 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

to classify the data sets and clustering method to cluster the classified as well as trained data sets. After the process noisy data's are removed from databases. Noisy example data sets are evaluated based on functional values. Analysis of outlier for noisy data are represented as follows

$Ot(f(x))=[P(f(X) \neq Y) \leq 2/m(d\ (ln(34em\ /d)log2(578m)) + ln(4/\delta)]$

Derived from the theorem: Consider F be a class of functions. With probability (w.p.) at least 1− δ over mid samples x from some fixed probability distribution P the following holds:

If f ∈ sign (F) has margin at least γ on the examples x then the expected error of f, d represent the base dimension of classifier.

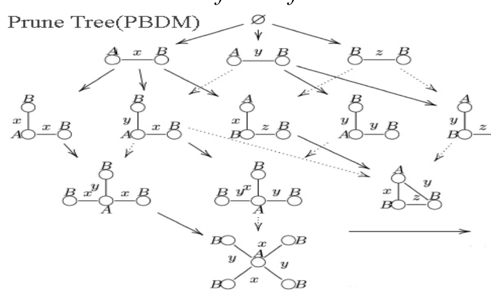

Figure 1: Prune Tree-OPBDM

4. KNOWLEDGE DISCOVERY FOR PRUNED DATA

Knowledge discovery from pruned data leads to unearthing of a great knowledge in data, domain understanding, selection processes, pre-processing of data and data elements, knowledge evaluation and consolidation and utilize the knowledge. The knowledge discovery process is interactive in nature and iterative in learning, decision making and for the decision support. It includes some basic steps for performing KDD process, they are listed as follows

- Data Selection
- Data Pre-processing
- Data Transformation
- Pattern evaluation
- Knowledge presentation

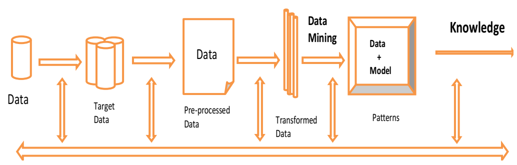

Figure 2: Knowledge Discovery Process

4.1 Data Selection

Relevant data to the analysis task are retrieved from the databases.

4.2 Data Pre-processing

To remove noisy and inconsistent data (i.e., cleaning and integration of multiple data sources).

4.3 Data Transformation

Data's are transformed into particular form appropriate for mining by performing various aggregation or composition operations.

4.4 Pattern evaluation

To identity the truly representing model for knowledge representation

4.5 Knowledge Presentation

Knowledge visualization and presentations are used to represent the mined knowledge

4.6 Prune Based Data Mining (OPBDM)

Pruning Decision tree focuses mainly on decision making and decision supportive models. Pruning an algorithms discard branches of trees that do not improve accuracy. To achieve this they implement one of two general paradigms: pre-pruning or post pruning [37] .Since the prune tree construction of various data sets for spatial databases using Prune based Data Mining algorithm can be achieved through basics of Post pruning, Pre pruning, building partial tree, Rule based tree construction set c. Here OPBDM algorithm reduces the complexity, Redundancy, Noisy Data's in spatial databases. Reduced-error pruning (post pruning) generates smaller and more accurate decision trees if pruning decisions are made using significance tests and the significance level is chosen appropriately for each dataset.

The attachment of a class label to each node of the tree is assumed to each node of the tree. For example, by taking the majority class of the training instances reaching that particular node, there are two classes namely X and Y. The tree predicted from figure 3 can be used to predict the test instances by filtering the whole tree to its leaf node which is corresponding of its instance attribute and class which assigns object class as well as class label to the particular leaf node. However using the Post pruning decision tree for potential classification over fits the training and instance data sets, randomly used in particular leaf along with its test and training set. Pre-pruning algorithms do not precisely implement "pruning" because they never prune live branches of a decision tree: they "prune" in advance by conquering the development of a branch if additional structure is not expected to raise accuracy [37].

Journal of Theoretical and Applied Information Technology
30th September 2014. Vol. 67 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

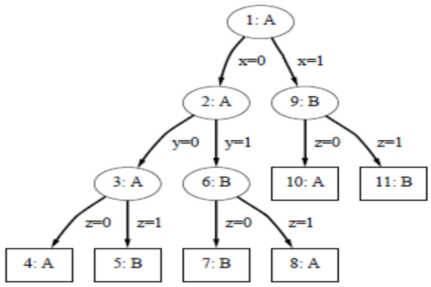

Figure 3: Post Pruning-Two Classes X and Y

Table 1: Pruned Data Set

X	Y	Class
0	0	A
1	0	B
0	1	B
1	1	A

Consider this as a hypothesis, *If tree A is pre-pruned using the parametric t-squared test, and tree B is pre-pruned using a corresponding permutation test, and both trees have the same size, then A will be less accurate than B.* when pre-pruning is applied, the significance test is used to ascertain the association between the values of an attribute and the class labels. At each node of the decision tree, pre-pruning methods apply a significance test to every attribute and class labels being considered for splitting. Splitting ceases if there is no significant attribute can be found. This means that several significance tests are performed at each node. Since multiple significance tests increases the probability that a significant association leads to more number of attributes being tested [37].

5. FRAMEWORK FOR OPBDM

Function PrunetreeBasedDataMining(X: attributes, I: instances, J: two class training Data, H: growing data, Q: pruning data, T: tree, P: pruning Data)
S: = attributes from X that are significant at level
R: =empty set of rules
If S is a empty return of values
X: = attribute which maximizes splitting criterion-object/attribute of S
for each value Vi of a
BuildaTree (A without b, instances from I with value Vi for b)
While tree of instance I is build
Set branch as no growth
End instance I
While not I empty
Divide I into rising set and pruning set
si := excellent rule for the positive class in the growing set
sr:= pruning of si with the top routine on the pruning setif performance of r is not enough then
r := empty rule assigning the negative class R: =put in sr to R
Delete the instances from I that are covered by sr;
Create a new instance for Expand Subset
Create a new instance for expanding data set Select a most excellent split using raising data H
divide t, H into subsets Hi of H and Q into subsets Qi of Q
While there are subsets Hi that are not yet been expanded and all subsets Hi expanded so far are leaf of H
Hm := subset with minimum entropy that has not been expanded
callExpandSubset(Hm,Qm)
f all the expanded subsets Gi are leaves
f := error rate for node on Q
f' := minimum error rate among all leaves according to Qi
If f ≤ f' then replace tree by leaf
If T is leaf node
return leaf node
for all branches of T
prune tree attached to branch
Check significance of each lean with respect to T's root
Remove all sub trees that do not have important extensions
e := error rate of T's root
e' := minimum error rate among all of T's significant leaf If no significant leaf can be found or e ≤ e'
replace tree by leaf
Return R
End;

6. REDUCING ERROR RATE IN PRUNING

- Evaluate each node for pruning
- Pruning is defined by removing the sub tree at that node, mark it as a leaf and allocate the most common class at that particular node
- When a node is removed, the resulting tree still performs the arbitrary action which is the same as original on the validation set, which removes the coincidence node and error nodes

Journal of Theoretical and Applied Information Technology
30th September 2014. Vol. 67 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Nodes are removed by performing proper iterations, selecting the node when it is removed from tree increases accuracy of the decision tree on the particular graph.
- Pruning is iterated until further pruning is risky, Pruning uses training and validation for test sets for large amount of data is available.

Prune Tree with n nodes is denoted as follow:

$Gx= [pf][\ \Gamma((n + 1)/2)\ (1 + x^{2} / n)^{(-n/2\ -\ 1/2)}\ ] / [\ \Gamma(n/2)\ \sqrt{}(P(\ln(n)))$

Here the G is the function to evaluate the error rate while we are performing tree pruning. P f(x(1)……x(n))denotes the number of node which are pruned. It performs iterative actions until we cannot perform further pruning to a tree. Table 2 to Table 4 show the performance measures of pruning. According to the experiment analysis, we shown that OPBDM algorithm reduces the complexity, Redundancy, Noisy Data's. Overall, the classification results of the OPDBM algorithm are acceptable when compare to ANN [3], C4.5 [36] and Regression analysis [39].

Table 2: Pruned Data Set

Sno	Pf	Gx
1	0.525	0.488
2	0.511	1.66
3	0.555	0.890
4	0.497	0.780
5	0.480	0.689
6	0.512	0.666
7	0.701	0.506

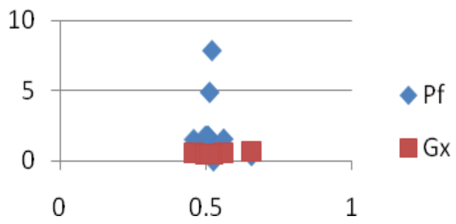

Figure 4: Pf & Gx Parameter Evaluation

Table 3 :Correlation Value Of Pf And Gx

Sno	P	Pf	Gx
1	0.455	0.525	0.488
2	0.560	0.511	1.66
3	0.652	0.555	0.890
4	0.540	0.497	0.780
5	0.551	0.480	0.689
6	0.435	0.512	0.666
7	0.350	0.701	0.506

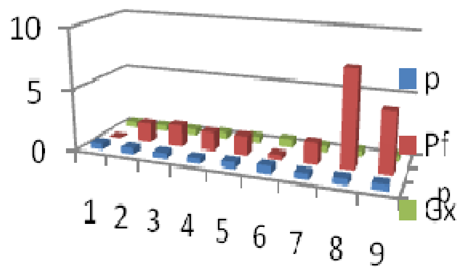

Figure 5: P, Pf & Gx Parameter Evaluation

Table 4:Correlation Value Of P,Pf And Gx

Sno	P	Pf
1	0.455	0.525
2	0.466	0.511
3	0.475	0.555
4	0.499	0.497
5	0.511	0.480
6	0.558	0.512
7	0.450	0.701
8	0.435	0.722
9	0.415	0.735
10	0.385	0.810
11	0.365	0.825
12	0.355	0.835
13	0.345	0.855
14	0.333	0.955
15	0.322	1.062
16	0.222	1.666
17	0.213	1.672

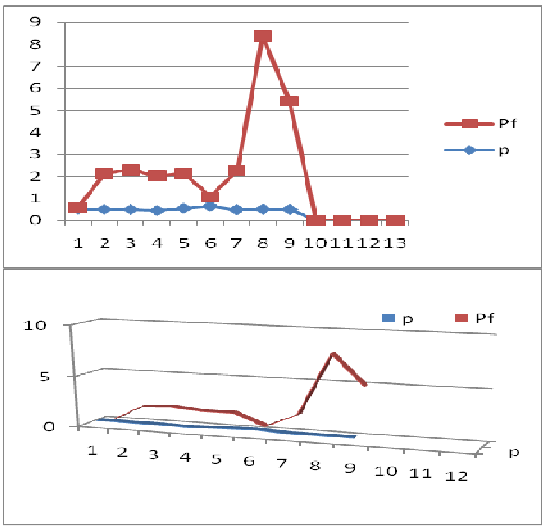

Figure 6 :Attribute Value Vs Class Value

Journal of Theoretical and Applied Information Technology
30th September 2014. Vol. 67 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

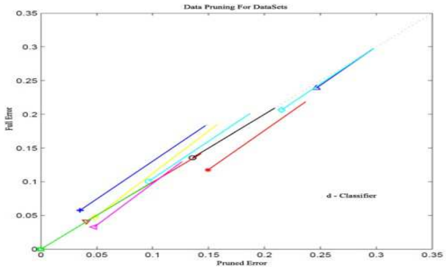

Figure 7:Reduced Prune Error Rate

7. CONCLUSION AND FUTURE ENHANCEMENT

This work presents a new foundational approach to prune based Data mining where a large number of datasets and records (multimedia based) are transformed in the spatial databases. Prune based decision support data mining methodologies to perform mining operations in spatial databases have been presented. The key challenge lies in preventing the data mining records from combining duplicates at different reliable levels to jointly restore the original data more perfect and accurate. The design forming part of this work has been proved, "the prune based Data mining" to have high dimensional accuracy in joint property, redundancy checks, Knowledge intelligence to gain knowledge with reliability in databases and data catalogues, then the databases will have no range in their joint restoration of the original data. In this paper verify the claim and demonstrate the effectiveness of our solution through numerical evaluation (figure 4 to figure 7) which stated as follows (described in the section 6).

$$Gx = [pf]\ [\ \Gamma((n + 1)/2)\ (1 + x^{2} / n)^{(-n/2 - 1/2)}\ ] / [\ \Gamma(n/2)\ \Gamma(P\ (\ln(n)))$$

This work takes initial step to enable prune based data mining with high perfection and accuracy. Many motivating and important instructions are rated by exploring it. For example, it is not clear how to expand the scope of other approaches in the area of partial information hiding, such as redundancy replacement, Domain specific datasets(for example PoS based). It is also of great interest to extend the approach in this paper to handle evolving Domain specific Datasets, handling multidimensional datasets with redundancy. Reduced-error pruning (post pruning) minimize their limitations and results are acceptable when compare to other classification techniques. In future research we will perform experimental analysis with various applications like sentiment mining for journal citation data using OPDBM algorithm.

REFERENCES:

[1] A. Agresti, *Categorical Data Analysis*. New York: Wiley, 1990.

[2] W.-H. Au and K. C. C. Chan, "Mining fuzzy association rules in a bank account database", *IEEE Trans. Fuzzy Syst.*, vol. 11, pp. 238– 248, Apr. 2003.

[3] C. Bishop, *Neural Networks for Pattern Recog.* New York: Oxford Univ. Press, 1995.

[4] B. P. Carlin and T. A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. London , U.K.: Chapman & Hall, 2000.

[5] K. C. C. Chan and W.-H. Au, "Mining fuzzy association rules in a database containing relational and transactional data", in *Data Mining and Computational Intelligence*, A. Kandel, M. Last, and H. Bunke, Eds. New York: Physica-Verlag, 2001, pp. 95–114.

[6] K. C. C. Chan and A. K. C.Wong, "APACS: A system for the automatic analysis and classification of conceptual patterns", *Comput. Intell.*, vol. 6, pp. 119–131, 1990.

[7] "A statistical technique for extracting classificatory knowledge from databases", in *KDD*, G. Piatetsky-Shapiro and W. J. Frawley, Eds. MenloPark, CA:/Cambridge, MA: AAAI/MIT Press, 1991, pp. 107–123.

[8] J. Y. Ching, A. K. C. Wong, and K. C. C. Chan, "Class-Dependent discretization for inductive learning from continuous and mixed-mode data", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 1–11, July 1995. J. Y. Ching, A. K. C. Wong, and K. C. C. Chan, "Class-Dependent discretization for

[9] A. Choenni, "Design and implementation of a genetic-based algorithm for data mining", in *Proc. 26th Int. Conf. Very Large Data Bases*, Cairo,Egypt, 2000, pp. 33–42.

[10] K. A. DeJong, W. M. Spears, and D. F. Gordon, "Using genetic algorithms for concept learning", *Mach. Learn.*, vol. 13, pp. 161–188, 1993.

[11] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview", in *Advances in KDDM*, U.M.Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Menlo Park, CA:/Cambridge: MA: AAAI/MIT Press, 1996, pp. 1–34.

[12] M. V. Fidelis, H. S. Lopes, and A. A. Freitas, "Discovering comprehensible classification

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

rules with a genetic algorithm", in *Proc.2000 Congress Evolutionary Computation*, San Diego, CA, 2000,pp. 805–810.

[13] D. B. Fogel, *Evolutionary Computation: Toward a New Philosophy ofMachine Intelligence*. Piscataway, NJ: IEEE Press, 1995.

[14] R.Forsyth, *PC/BEAGLE User's Guide*. Nottingham, U.K.: Pathway Research Ltd., 1990.

[15] A. A. Freitas, "Understanding the critical role of attribute interaction in data mining", *Artif. Intell. Rev.*, vol. 16, pp. 177–199, 2002.

[16] J. Gehrke, V. Ganti, R. Ramakrishnan and W.-Y. Loh, "BOAT – Optimistic decision tree construction", in *Proc. ACM SIGMOD Int. Conf.Management of Data*, Philadelphia, PA, 1999, pp. 169–180.

[17] J. Gehrke, R. Ramakrishnan, and V. Ganti, "RainForest – A frameworkfor fast decision tree construction of large datasets", in *Proc. 24th Int.Conf. Very Large Data Bases*, New York, 1998, pp. 416–427.

[18] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

[19] D. P. Greene and S. F. Smith, "Using coverage as a model building constraint in learning classifier systems", *Evol. Comput.*, vol. 2, no. 1, pp.67–91, 1994.

[20] R. R. Hill, "A Monte Carlo study of genetic algorithm initial population generation methods", in *Proc. 31st Conf.Winter Simulation–A Bridge to the Future*, Phoenix, AZ, 1999, pp. 543–547.

[21] J. Holland, "Escaping brittleness: The possibilities of general-purpose learning algorithms applied to parallel rule-based systems in Machine Learning: An Artificial Intelligence Approach*"*, R. Michalski, J. Carbonell, and T. Mitchell, Eds. San Mateo, CA: Morgan Kaufmann, 1986.

[22] H. Ishibuchi and T. Nakashima, "Improving the performance of fuzzy classifier systems for pattern classification problems with continuous attributes", *IEEE Trans. Ind. Electron.*, vol. 46, pp. 1057–1068, Dec. 1999.

[23] C. Z. Janikow, "A knowledge-intensive genetic algorithm for supervised learning", *Mach. Learn.*, vol. 13, pp. 189–228, 1993.

[24] B. A. Julstrom, "Seeding the population: Improved performance in a genetic algorithm for the rectilinear steiner problem", in *Proc. ACMSymp. Applied Computing*, Phoenix,AZ, 1994, pp. 222–226.

[25] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision tree hybrid", in *Proc. 2nd Int. Conf. KDDM*, Portland, OR, 1996.

[26] W. Kwedlo and M. Kretowski, "Discovery of decision rules from databases: An evolutionary approach", in *Proc. 2nd European Symp. Principles of Data Mining and Knowledge Discovery*, Nantes, France, 1998, pp. 370–378.

[27] R. J. Light and B. H. Margolin, "An analysis of variance for categorical data", *J. Amer. Statist. Assoc.*, vol. 66, pp. 534–544, 1971.

[28] J. Lockwood, "Study predicts 'Epidemic' churn", in *Wireless Week*, Aug. 25, 1997.

[29] A. D. McAulay and J. C. Oh, "Improving learning of genetic rule-based classifier systems", *IEEE Trans. Syst. Man, Cybern.*, vol. 24, pp.152–159, Jan. 1994.

[30] M. Mehta, R. Agrawal, and J. Rissanen, "SLIQ: A fast scalable classifier for data mining", in *Proc. 5th Int. Conf. Extending Database Technology*, Avignon, France, 1996, pp. 18–32.

[31] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, 3rd Revised and Extended New York: Springer-Verlag, 1996.

[32] M. C. Mozer, R. Wolniewicz, D. B. Grimes, E. Johnson, and H.Kaushansky, "Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry", *IEEE Trans.Neural Networks*, vol. 11, pp. 690–696, May 2000.

[33] M. O. Noordewier, G. G. Towell, and J. W. Shavlik, "Training knowledge-based neural networks to recognize genes in DNA sequences", in *Advances in Neural Information Processing Systems*, R.P. Lippmann,J.E. Moody, and D.S. Touretzky, Eds. San Mateo, CA: Morgan Kaufmann, 1991, vol. 3.

[34] J. R. Quinlan, "Decision trees as probabilistic classifiers", in *Proc. 4th Int. Workshop Machine Learning*, Irvine, CA, 1987, pp. 31–37.

[35] "Simplifying decision trees", *Int. J. Man-Mach. Stud.*, vol. 27, pp. 221–234, 1987.

[36] *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[37] R. Rastogi and K. Shim, "PUBLIC: A decision tree classifier that integrates building and pruning", in *Proc. 24th Int. Conf. Very Large Databases*, New York, 1998, pp.404-415.

[38] J. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A scalable parallel classifier for data mining", in *Proc. 22nd Int. Conf. Very Large*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Data Bases, Mumbai (Bombay), India, 1996, pp. 544–555.

[39] S. Smith, "Flexible learning of problem solving heuristics through adaptive search", in *Proc. 8th Int. Joint Conf. Artificial Intelligence*, Karlsruhe, Germany, 1983, pp.422–425

[40] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S.Johannes, "Usingthe ADAP learning algorithm to forecast the onset of diabetes mellitus", *Proc. Symp. Computer Applications and Medical Cares*, pp. 261–265, 1988.

[41] C. H.Yang, "The effects of initial population in genetic search for time constrained traveling salesman problems", in *Proc. ACM Conf. Computer Science*, Indianapolis, IN, 1993, pp. 378–383.

[42] Autonomous Decision-Making: A Data Mining Approach, Andrew Kusiak, Jeffrey A. Kern, Kemp H. Kernstine, and Bill T. L. Tseng. IEEE transactions on information technology in biomedicine, vol. 4, NO. 4, DEC 2000.

[43] A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction, Wai-Ho Au, Keith C. C. Chan, and Xin Yao*, Fellow,* IEEE transactions on evolutionary computation, vol. 7, no. 6, december 2003.

[44] Rough set theory analysis, www.nit*.eu/czasopisma/JTIT/2002/3/7.pdf*

[45] Enhancing Concept Detection by Pruning Data with MCA-based Transaction Weights, Lin Lin, Mei-Ling Shyu, Shu-Ching Chen, *Dept. of ECE, University of MiamiCoral Gables, FL 33124,FL 33199, USA*