International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 4, Jul-Aug 2014

RESEARCH ARTICLE OPEN ACCESS

Design and Development of Data Mining System to Analyze Cars using Improved ID3 with T*k*NN Clustering Algorithm

M.Jayakameswaraiah[1], Prof.S.Ramakrishna[2]
Ph.D. Research Scholar[1], Professor[2],
Department of Computer Science,
Sri Venkateswara University, Tirupati,
AP-India

ABSTRACT

Conventional way of business is a challenging in car market due to many competitors are there around the world for providing competitive products. The car manufacturers categorizes the car users and have to invent a suitable car; the seller correctly groups the buyers and he sells a right car; and the customers selects best car by analyzing more brands of cars with 'N' number of sellers. These three cases they spent too much of time for analyzing old or statistical data for choosing a right product. Now a day's customers are required comfort and their loving brand & color. With the advent of the Internet and Data Mining Algorithms has undoubtedly contributed to the shift of marketing focus. In this paper, we proposed Improved ID3 with T*k*NN algorithm for best car market analysis. We have executed the same in WEKA Tool with Java code. We analyzed the graphical performance analysis between T*k*NN and our novel improved ID3 with T*k*NN clustering algorithms with Classes to Clusters evaluation purchase, safety, luggage booting, persons (seating capacity), doors, maintenance and buying attributes of customer's requirements for unacceptable/acceptable/good/very good ratings of a car to purchase.

Keywords: - ID3 Algorithm, KNN, Improved ID3 with T*k*NN Algorithm

I. INTRODUCTION

Economic growth of a country depends on transportation as one constraint. Like many economic activities that are intensive in the use of infrastructures, the transport sector is an important component of the economy impacting on development and the welfare of population. A relation between the quantity and quality of transport infrastructure and the level of economic development is apparent. When transport systems are efficient, they provide economic and social opportunities and benefits that result in positive multipliers effects such as better accessibility to markets, employment and additional investments.

A new business culture is developing today [4, 7, 24, 19]. Within it, the economics of customer relationships are changing in fundamental ways, and companies are facing the need to implement new solutions and strategies that address these changes. The concepts of mass production and mass marketing, first created during the Industrial Revolution, are being supplanted by new ideas in which customer relationships are the central business issue [1, 5]. Firms today are concerned with increasing customer value through analysis of the customer lifecycle. The tools and technologies of data warehousing, data mining, and other customer relationship techniques afford new opportunities for businesses to act on the concepts of relationship marketing. The old model of "design-build-sell" (a product-oriented view) is being replaced by "sell-build-redesign" (a customer-oriented view). It is a spiral model of software engineering [26]. The traditional process of mass marketing is being challenged by the new approach of one-to-one marketing. In the traditional process, the marketing goal is to reach more customers and expand the customer base [31, 11, 5]. But given the high cost of acquiring new customers, it makes better sense to conduct business with current customers. In so doing, the marketing focus shifts away from the breadth of customer base to the depth of each customer's needs.

The lifecycle of a modern car comprises a multitude of complex and interdependent tasks that start early during the development phase, guide and advice the production process and keep track of issues related to operating vehicles. We will present examples of data mining applications from all these three stages: development, production planning and fault analysis. All contributions share the property that we use (or extract) rule patterns to explain the domain under analysis to the user. Rules (in form of association rules) are a well-understood means of representing knowledge and data dependencies. The inherent interdisciplinary character of the automobile development and manufacturing process requires models that are easily understood across application area boundaries. The understanding of patterns can be greatly enhanced by providing powerful visualization methods alongside with the analysis tools.

The next section will briefly sketch the underlying theoretical frameworks, after which we will present and discuss successfully applied fault analysis, planning and development methods, all of which have been rolled out to production sites of two large automobile manufacturers.

Section 2 provides some notations needed for the rest of the article. Section 3 discusses an approach based on graphical models to assess and reveal potential fault patterns inside vehicle data. Section 4 deals with the handling of production planning. Section 5 discovers rules from time series that are created by prototype simulations. Finally, section 6 concludes the findings.

II. BACKGROUND

Selecting the suitable car is extremely tricky job if parameters (color, comfort, seating capacity, maintenance, price, and so on) are known otherwise it is difficult task. If the customer knows these all things then also sometimes it is hard to choose the right car.

The problem is unmanageable in the perspective of manufacturer and seller, because they must work with different categories of people [31]. Some people preferred only high cost cars, some are low price with all features and others are in between these classes. One more category of people are only knowing information about different brands but they never buys.

The need to increase the productivity of manufacturer, raises the seller transactions and customer satisfy of the selected car comforts. Comfort transportation is encourages frequency of vehicle usage, it increases Economic growth as well as it decreases wastage of time in journey. Car is the symbol of comfortable.

If this system is available then there are no capabilities required to assist with telecom expense management, i.e., the administrator can find out the number of calls and text messages used as well as cellular and WiFi data usage, both for home and roaming networks [19].

We will now briefly discuss the notational underpinning that is needed to present the ideas and results from the industrial applications.

Graphical Models

As we have pointed out in the introduction, there are dependencies and independencies that have to be taken into account when reasoning in complex domains shall be successful. Graphical models are appealing since they provide a framework of modeling independencies between attributes and influence variables. The term “graphical model” is derived from an analogy between stochastic independence and node separation in graphs. Let V = {A1 ,... , An } be a set of random variables. If the underlying probability distribution P (V) satisfies some criteria (see e. g. (CGH97; Pea93)), then it is possible to capture some of the independence relations between the variables in V using a graph G = (V, E), where E denotes the set of edges. The underlying idea is to decompose the joint distribution P (V) into lower-dimensional marginal or conditional distributions from which the original distribution can be reconstructed with no or at least as few errors as possible (LS88; Pea88). The named independence relations allow for a simplification of these factor distributions. We claim, that every independence that can be read from a graph also holds in the corresponding joint distribution. The graph is then called an independence map (see e. g. (BSK09)).

Association Rules

The introduction of frequent item set mining and subsequently association rule induction (AIS93; AMS+ 96) has created a prospering field of data mining. It is the simplicity of the underlying concept that allowed for a broad acceptance among all kinds of users no matter whether they possess a data analysis background or not. An association rule is basically an if-then rule. The if -part is called antecedent while the then -part is named the consequent. Both may consist of conjunctions of attribute-value pairs, however, the consequent often consists of only one pair. An example of an association rule could be

If a person is male and a smoker, his probability of having lung cancer is 10%.

This corresponds to the imagination that we pick a person at random from an underlying population (the database) and observe its properties, which is its attribute values. The above rule can then be represented in a more formal fashion as

$$\text{Gender} = \text{male} \wedge \text{Smoker} = y \rightarrow \text{Cancer} = y$$

We refer to a database case as being covered by a rule if the antecedent and consequent attributes values match. For instance, a smoking man having lung cancer would be covered by the above rule. The general form of a rule has the following form:

$$A_1 = a_1 \bigwedge \dots \bigwedge A_n = a_n \rightarrow C = c \overset{\text{abbr}}{=} a \rightarrow c$$

We will only discuss rules with one consequent attribute which will be a class variable. We thus use the notions class and consequent interchangeably.

Since not every database entry matching the antecedent also matches the consequent it is necessary to record this information. The probability that a database case matching the antecedent also matches the consequent, that is P (c | a), is called the confidence of the rule. The above rule 1 has a confidence of 0.1. There is a multitude of other measures that quantify certain aspects of a rule. We will briefly discuss those that are used in this paper.

The number of cases covered by the rule is referred to as the (absolute) support of the rule. The relative support equals P (a, c); it is the absolute support divided by the database size. The recall quantifies the fraction (or probability if you keep the above scenario of picking at random) of database cases matching the antecedent, given the consequent. In other words: What is the probability of a person being male and a smoker if this person has cancer? As a last measure (the only unbounded one) we introduce the lift. It represents the ratio between the confidence P (c | a) and the marginal consequent probability P (c): Let the marginal cancer rate be 0.01. Then, rule 1 has a lift of 10 since the confidence is ten times larger than the marginal cancer rate. We summarize the measures below:

- relative support: rel-supp(a → c) = P (a, c)
- confidence: conf(a → c) = P (c | a)

- recall: recall(a → c) = P (a | c)
- lift: lift(a → c) = P (c | a)

About WEKA tool

There are many tools available for data mining and machine learning, but in this thesis we use the open source software suite WEKA which stands for Waikato Environment for Knowledge Analysis. The main reason why we selected to use WEKA was because of its versatility. WEKA is a popular tool used for data analysis, machine learning and predictive modeling that was developed by the University of Waikato in New Zealand using the programming language JAVA.

Main Features

Some of WEKAs main features are the following:

Data preprocessing - WEKA supports a couple of popular text file formats such as CSV, JSON and Matlab ASCII files to import data along with their own file format ARFF. They also have support to import data from databases through JDBC. Besides importing data, they have a wide collection of supervised as well as unsupervised filters to apply on your data to facilitate further analysis.

Data classification - A huge collection of algorithms have been implemented to perform classification on data sets. These include Bayesian algorithms, mathematical functions such as support vector machines, lazy classifiers implementing nearest-neighbor calculations; Meta based algorithms as well as rule and tree-based classifiers.

Data clustering - A couple of algorithms for clustering exist such as variations of the k-mean method as well as density and hierarchical based clustering algorithms.

Attribute association - Methods to analyze data using association rule learners. Association rules can be seen as rules describing relations between attributes in a data set.

Attribute selection - Methods to evaluate which attribute contribute the most when predicting an outcome.

Data visualization - Depending on the methods used to analyze the data, this view can to plot data against suitable variables as well as give tools to analyze specific points further.

III. ID3 ALGORITHM

The ID3 algorithm was originally developed by J. Ross Quinlan at the University of Sydney, and he first presented it in the 1975 book "Machine Learning". The ID3 algorithm induces classification models, or decision trees, from data. It is a supervised learning algorithm that is trained by examples for different classes. After being trained, the algorithm should be able to predict the class of a new item.

ID3 identifies attributes that differentiate one class from another. All attributes must be known in advance, and must also be either continuous or selected from a set of known values. For instance, temperature (continuous), and country of citizenship (set of known values) are valid attributes. To determine which attributes are the most important, ID3 uses the statistical property of entropy. Entropy measures the amount of information in an attribute. This is how the decision tree, which will be used in testing future cases, is built.

The principle of the ID3 algorithm is as follows. The tree is constructed top-down in a recursive fashion. At the root, each attribute is tested to determine how well it alone classifies the transactions. The "best" attribute (to be discussed below) is then chosen and the remaining transactions are partitioned by it.

Entropy

In information theory, entropy is a measure of the uncertainty about a source of messages. The more uncertain a receiver is about a source of messages, the more information that receiver will need in order to know what message has been sent.

Information gain

Now consider what happens if we partition the set on the basis of an input attribute X into subsets $T_1, T_2, T_3, \dots, T_N$. The information needed to identify the class of an element of T is the weighted average of the information needed to identify the class of an element of each subset:

$$H(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i)$$

In the context of building a decision tree, we are interested in how much information about the output attribute can be gained by knowing the value of an input attribute . This is just the difference between the information needed to classify an element of before knowing the value of X, H(T), and the information needed after partitioning the dataset T on the basis of knowing the value of X, H(X, T). We define the information gain due to attribute X for set T as:

$$\text{Gain}(X,T) = H(T) - H(X,T)$$

In order to decide which attribute to split upon, the ID3 algorithm computes the information gain for each attribute, and selects the one with the highest gain.

The simple ID3 algorithm above can have difficulties when an input attribute has many possible values, because Gain(X, T) tends to favor attributes which have a large number of values. It is easy to understand why if we consider an extreme case.

Imagine that our dataset contains an attribute that has a different value for every element of T. This could arise in practice if a unique record ID was retained when extracting, from a database.

The problem also arises when an attribute can take on many values, even if they are not unique to each element. Quinlan (1986) suggests a solution based on considering the amount of information required to determine the value of an attribute X for a set T. This is given by H(PX,T), where PX,T is the probability distribution of the values of X:

$$P_{X,T} = \left(\frac{|T_1|}{|T|}, \frac{|T_2|}{|T|}, \frac{|T_3|}{|T|}, \dots, \frac{|T_n|}{|T|} \right)$$

The quantity H(PX,T) is known as the split information for attribute X and set T.

ID3 (Learning Sets S, Attributes Sets A, Attributesvalues V)
Begin

- Load training data set for training.
- If all examples are positive, return the single-node tree root with label is positive.
- If all examples are negative, return the single-node tree root with label is negative.
- If number predicting attributes is empty, then return the single node tree root, with the label is most common value of the target attribute in the examples.
- Otherwise
- Begin
  For rootNode, we compute Entropy(rootNode.subset) first

$$\text{Entropy (S)} = \sum_{i=1}^{c} P_i \log_2 P_i$$

  - ✓ If Entropy(rootNode.subset)==0, then rootNode.subset consists of records all with the same value for the categorical attribute, return a leaf node with decision attribute:attribute value;
  - ✓ If Entropy(rootNode.subset)!=0, then compute information gain for each attribute left(have not been used in splitting), find attribute A with Maximum(Gain(S,A)). Create child nodes of this rootNode and add to rootNode in the decision tree.
  - ✓ For each child of the rootNode, apply ID3(S,A,V) recursively until reach node that has entropy=0 or reach leaf node.
- End

End

IV. KNN ALGORITHM

The k-nearest neighbor (KNN) algorithm is a simple and one of the most intuitive machine learning algorithms that belongs to the category of instance-based learners. Instance-based learners are also called lazy learner because the actual generalization process is delayed until classification is performed, i. e., there is no model building process. Unlike most other classification algorithms, instance-based learners do not abstract any information from the training data during the learning (or training) phase. Learning (training) is merely a question of encapsulating the training data, the process of generalization beyond the training data is postponed until the classification process.

The high degree of local sensitivity makes kNN highly susceptible to noise in the training data – thus, the value of k strongly influences the performance of the kNN algorithm. The optimal choice of k is a problem dependent issue, but techniques like cross-validation can be used to reveal the optimal value of k for objects within the training set.

General evaluation considering the simplicity of the KNN algorithm, the classification results of KNN are generally quite good and comparable to the performance achieved with decision trees and rule-based learners. However, the class specification accuracy of KNN models does in general not reach the accuracy achieved with support vector machines or ensemble learners. KNN is considered to be intolerant to noise, since its similarity measures can easily be distorted by errors in the attribute values, and is also very sensitive to irrelevant features. On the contrary, KNN models are usually not prone to over fitting and can be applied to incremental learning strategies – since KNN does not build a classification model, newly classified instances can be added to the training set easily.

There are several studies that survey the application of KNN for classification tasks. Besides almost all introductory data mining books and surveys, that summarizes several improvements of KNN algorithms for classification. Distance tables are calculated to produce real-valued distances from features coming from symbolic domains. It is standard KNN in three different application domains and has advantages in training speed and simplicity. On a weight-adjusted KNN implementation which finds the optimal weight vector using an optimization function based on the leave-out-out cross-validation and a greedy hill climbing technique.

The k-NN search is conducted in two phases. A Z-order-based approximate proximity mea- sure is used to find the approximate k-NN. Next, a recursive correction algorithm is used to improve the accuracy. Another set of techniques is based on a hybrid of spatial subdivision up to a threshold granularity and small scale brute force evaluation or heuristics for refinement. Some techniques take advantage of the intrinsic dimensionality of the data set to project the data set into a low dimensional space that preserves proximity.

kNN Algorithm (Set startAndEndPoint, real *ε* , int MinC)
Begin
Compute , the distance between z and every object, .
select , the set of k closet training objects to z.
End

V. IMPROVED ID3 WITH TKNN CLUSTERING ALGORITHM

This Algorithm is characterized by the ability to deal with the explosion of business data and accelerated market changes, these characteristics help providing powerful tools for decision makers, such tools can be used by business users (not only statisticians) for analyzing huge amount of data for patterns and trends [11]. Consequently, data mining has become a research area with increasing importance and it involved in determining useful patterns from collected data or determining a model that fits best on the collected data.

It is used to investigate the attributes of car in the perspective of manufacturer, seller and customer. It is essential to analyze the car in short span of time, consider cases when all parties (i.e. manufacturer, seller and customer) selecting a right product.

ImprovedID3WithT*k*NN (Learning Sets S, Attributes Sets A, Attributesvalues V, Y£)

Begin

1. Load training data set for training.
2. If attributes are uniquely identified in data set, remove it from training set.
3. On the basis of distance metric divide the given training data into subsets.
   - 3.1. Calculate the distance for n objects, each instance in available dataset.

$$D(x,y) = \left[\sum_{i=1}^{n} |X_i - Y_i|\right]$$

   Where X is selected instance and Y is comparing instance.

4. if D>55% then instance is belong to same group and add into new set and remove from original data set. Otherwise do nothing.
5. Repeat the steps 3.1 and 4 for each instance until all matched it not found.
6. On each subset apply ID3 algorithm recursively.
   - ➢ If all examples are positive, return the single-node tree root with label is positive.
   - ➢ If all examples are negative, return the single-node tree root with label is negative.
   - ➢ If number predicting attributes is empty, then return the single node tree root, with the label is most common value of the target attribute in the examples.
   - ➢ Otherwise

     Begin
     - ✓ For rootNode, we compute Entropy(rootNode.subset) first

$$\text{Entropy }(S) = \sum_{i=1}^{c} P_i \log_2 P_i$$

     - ✓ If Entropy(rootNode.subset)==0, then rootNode.subset consists of records all with the same value for the categorical attribute, return a leaf node with decision attribute:attribute value;
     - ✓ If Entropy(rootNode.subset)!=0, then compute information gain for each attribute left(have not been used in splitting), find attribute A with Maximum(Gain(S,A)). Create child nodes of this rootNode and add to rootNode in the decision tree.
     - ✓ For each child of the rootNode, apply ID3(S,A,V) recursively until reach node that has entropy=0 or reach leaf node.

     End

7. Construct T*k*NN graph among instances.
8. Initialize the similarities on each edge as $W_{iz} = \exp\left(\frac{\| x_i - x_z \|^2}{2\sigma^2}\right)$ and normalize to $\sum_z W_{iz} = 1$.
9. Determine the α_u^j values for all unlabeled data.
10. Compute the label set prediction matrix P.
11. Predict label set for each unbalanced instance by $y_i = \text{Sign}(P\alpha_i)$ $(\forall i \in u)$

End

We have executed the same in Weka Tool with Java code and compared the performance of two algorithms based on different Percentage Splits to help the car seller/manufacturer for analyzing their customer views in purchasing a car.

We analyzed the graphical performance analysis between KNN and our novel improved ID3 with T*k*NN clustering algorithms with Classes to Clusters evaluation purchase, safety, luggage booting, persons (seating capacity), doors, maintenance and buying attributes of customer's requirements for unacceptable/acceptable/good/very good ratings of a car to purchase.

VI. RESULT ANALYSIS

Improved ID3 Algorithm with T*k*NN using Percentage Split 66%

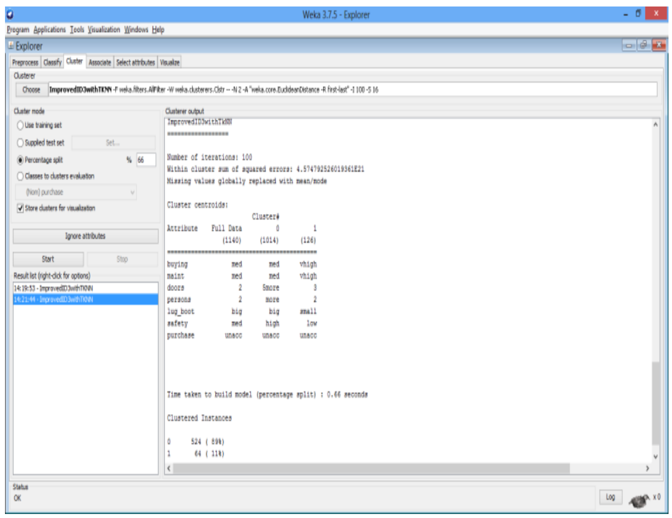

Figure 1: Test Result on Car Data Set with Percentage Split of 66% using Improved ID3 Algorithm with TkNN

Improved ID3 Algorithm with T*k*NN Using Training Set

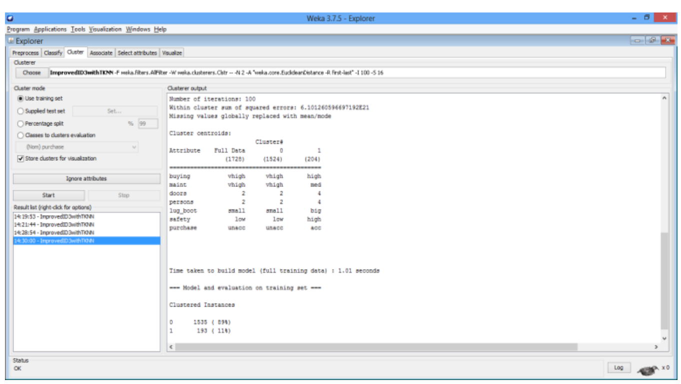

Figure 2: Test Result on Car Data Set with Training Set using Improved ID3 Algorithm with TkNN

Improved ID3 Algorithm with TkNN using Classes to Clusters evaluation safety attribute

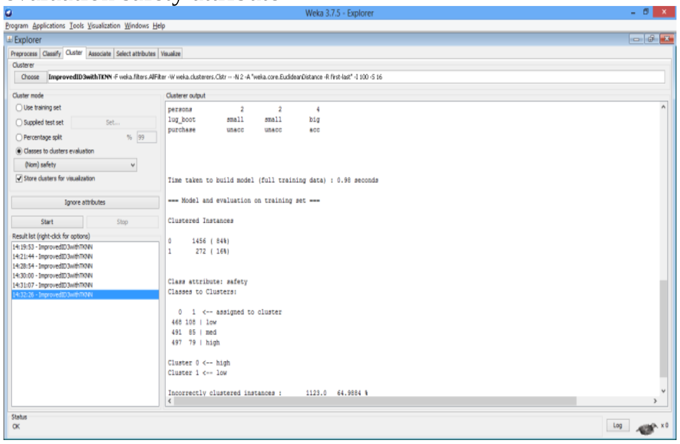

Figure 3: Test Result on Car Data Set with Classes to Clusters evaluation safety attribute

1) Visualize Curve

Improved ID3 Algorithm with TkNN using Training Set

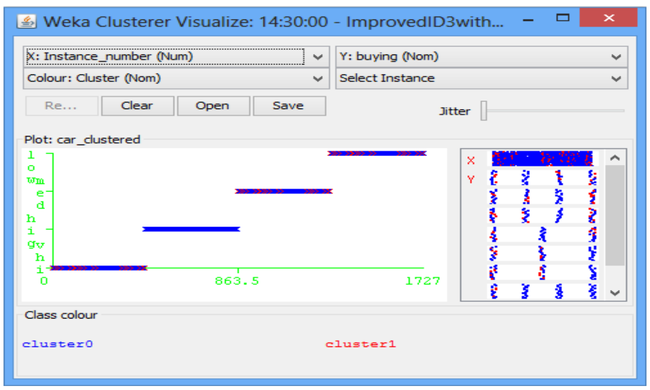

Figure 4: Visualize cluster assignments for Car Data Set using Improved ID3 Algorithm with TkNN

2) Within cluster sum of squared errors

This error value given cluster is computed by: for each instance in the cluster, summing the squared differences between each attributes value and the corresponding one in the cluster centroid. These are summed up for each instance in the cluster and for all clusters.

The within cluster sum of squared errors are measured on all the training data, so selecting the best result that you get is not necessarily going to be the best for future data due to possible over fitting.

Within cluster sum of squared errors	Improved ID3 Algorithm with TkNN
Percentage split 33%	2.319678065
Percentage split 66%	4.574792526
Percentage split 99%	6.871412161
Training Set	6.101260597
Classes to Clusters evaluation purchase attribute	6.013638562
Classes to Clusters evaluation safety attribute	5.736937402
Classes to Clusters evaluation luggage booting attribute	5.257322056
Classes to Clusters evaluation persons attribute	5.704655599
Classes to Clusters evaluation doors attribute	5.469459613
Classes to Clusters evaluation maintenance attribute	5.561693333
Classes to Clusters evaluation buying attribute	5.575528391

Table 1: Improved ID3 Algorithm with TkNN with Sum of within cluster distances

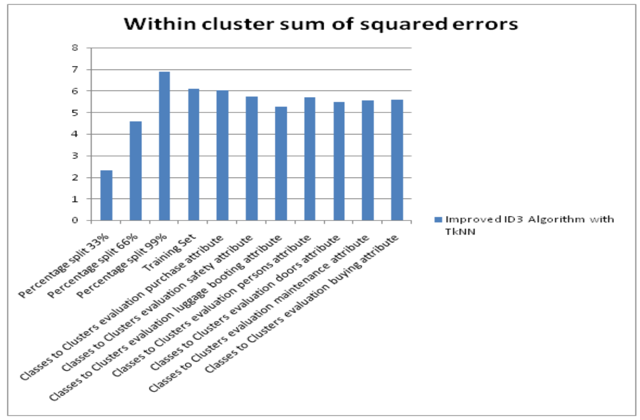

Figure 5: Improved ID3 Algorithm with TkNN with Sum of within cluster distances

3) Number of Iterations

This parameter explains one classifier with training data and tested against test data with these many specified number of times.

Number of iterations	Improved ID3 Algorithm with TkNN
Percentage split 33%	100
Percentage split 66%	100
Percentage split 99%	100
Training Set	100
Classes to Clusters evaluation purchase attribute	100
Classes to Clusters evaluation safety attribute	100
Classes to Clusters evaluation luggage booting attribute	100
Classes to Clusters evaluation persons attribute	100
Classes to Clusters evaluation doors attribute	100
Classes to Clusters evaluation maintenance attribute	100
Classes to Clusters evaluation buying attribute	100

Table 2: Improved ID3 Algorithm with TkNN with Number of Iterations

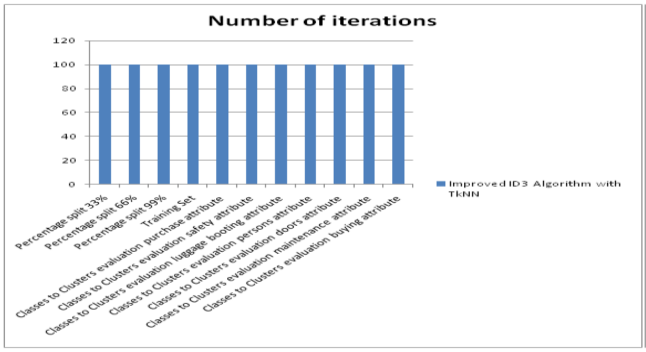

Figure 6: Improved ID3 Algorithm with TkNN with Number of Iterations

4) Cluster Instances of TkNN Algorithm

attribute	Total(1728)	Clster0(1062)	Cluster1(666)
buying	vhigh	low	vhigh
maint	vhigh	vhigh	high
doors	2	5more	2
persons	2	more	2
lug_boot	small	small	small
safety	low	low	hig
purchase	unacc	unacc	unacc

Table 3: Buying attribute Comparison on TkNN and ImprovedID3withTkNN

5) Cluster Instances of Improved ID3 Algorithm with TkNN Algorithms

attribute	Total(1728)	Clster0(1524)	Cluster1(204)
buying	vhigh	vhigh	high
maint	vhigh	vhigh	med
doors	2	2	4
persons	2	2	4
lug_boot	small	small	big
safety	low	low	high
purchase	unacc	unacc	acc

Table 4: Buying attribute Comparison on TkNN and ImprovedID3withTkNN

A. Buying attribute Comparison on TkNN and ImprovedID3withTkNN

	TkNN (cluster 0)	**Improved ID3 with TkNN (cluster 0)**	**TkNN (cluster 1)**	**Improved ID3 with TkNN (cluster 1)**
Very High		1524	666	
High				204
Medium				204
Low	1062			

Table 5: Buying attribute Comparison on TkNN and ImprovedID3withTkNN

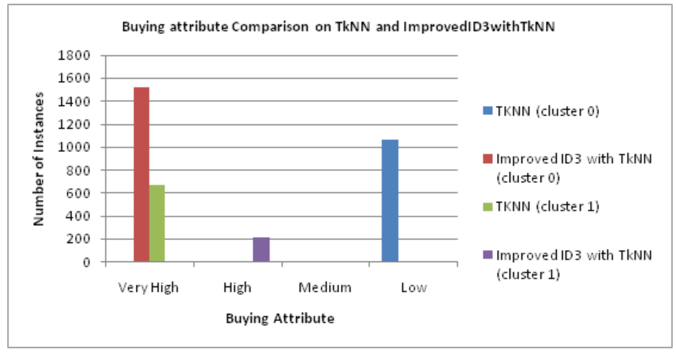

Figure 7: Buying attribute Comparison on TkNN and ImprovedID3withTkNN

B. Maintenance attribute Comparison on TkNN and ImprovedID3withTkNN

	TkNN (cluster 0)	**Improved ID3 with TkNN (cluster 0)**	**TkNN (cluster 1)**	**Improved ID3 with TkNN (cluster 1)**
Very High	1062	1524		
High			666	
Medium				204
Low				

Table 6: Maintenance attribute Comparison on TkNN and ImprovedID3withTkNN

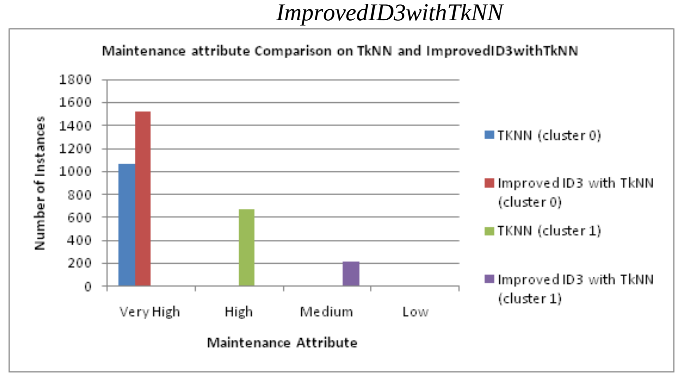

Figure 8: Maintenance attribute Comparison on TkNN and ImprovedID3withTkNN

C. Doors attribute Comparison on TkNN and ImprovedID3withTkNN

	TkNN (cluster 0)	**Improved ID3 with TkNN (cluster 0)**	**TkNN (cluster 1)**	**Improved ID3 with TkNN (cluster 1)**
5 or More Doors	1062			
4 Doors				204
3 Doors				
2 Doors		1524	666	

Table 7: Doors attribute Comparison on TkNN and ImprovedID3withTkNN

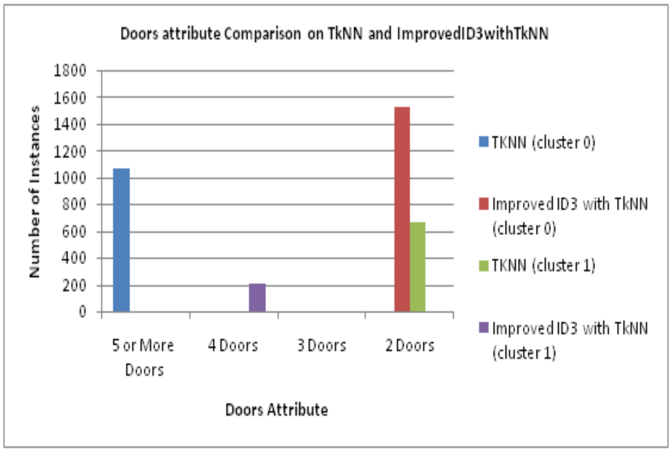

Figure 9: Doors attribute Comparison on TkNN and ImprovedID3withTkNN

D. Persons attribute Comparison on TkNN and ImprovedID3withTkNN

	T*k*NN (cluster 0)	**Improved ID3 with T*k*NN (cluster 0)**	**T*k*NN (cluster 1)**	**Improved ID3 with T*k*NN (cluster 1)**
More Persons	1062			204
4 Persons				
2 Persons		1524	666	

Table 8: Persons attribute Comparison on TkNN and ImprovedID3withTkNN

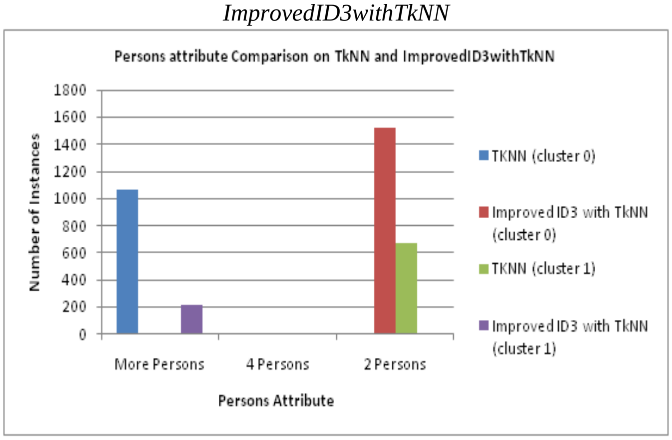

Figure 10: Pesons attribute Comparison on TkNN and ImprovedID3withTkNN

E. Luggage Booting attribute Comparison on TkNN and ImprovedID3withTkNN

	T*k*NN (cluster 0)	**Improved ID3 with T*k*NN (cluster 0)**	**T*k*NN (cluster 1)**	**Improved ID3 with T*k*NN (cluster 1)**
Big				204
Medium				
Small	1062	1524	666	
very high				

Table 9: Luggage Booting attribute Comparison on TkNN and ImprovedID3withTkNN

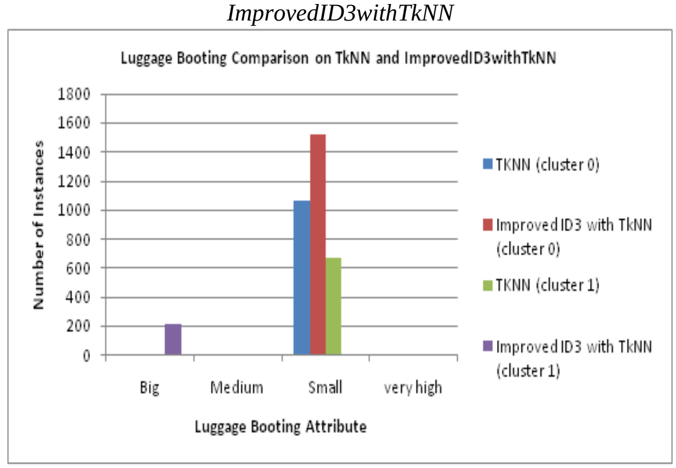

Figure 11: Luggage Booting attribute Comparison on TkNN and ImprovedID3withTkNN

F. Safety attribute Comparison on TkNN and ImprovedID3withTkNN

	T*k*NN (cluster 0)	**Improved ID3 with T*k*NN (cluster 0)**	**T*k*NN (cluster 1)**	**Improved ID3 with T*k*NN (cluster 1)**
High			666	204
Medium				
Low	1062	1524		

Table 10: Safety attribute Comparison on TkNN and ImprovedID3withTkNN

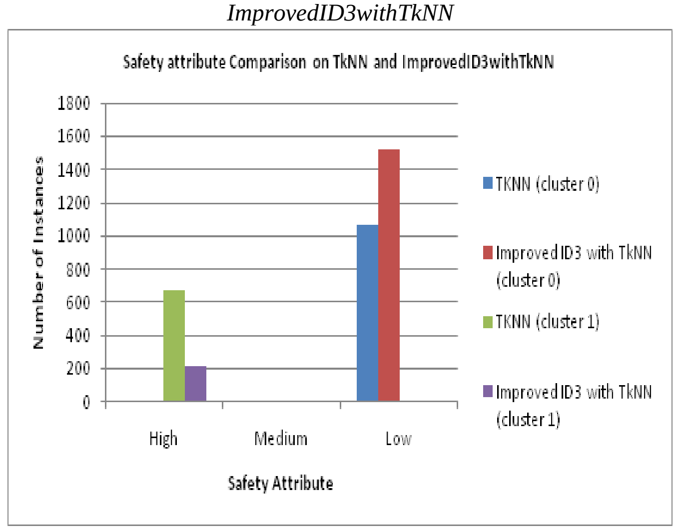

Figure 12: Safety attribute Comparison on TkNN and ImprovedID3withTkNN

G. Purchase attribute Comparison on TkNN and ImprovedID3withTkNN

	T*k*NN (cluster 0)	**Improved ID3 with T*k*NN (cluster 0)**	**T*k*NN (cluster 1)**	**Improved ID3 with T*k*NN (cluster 1)**
Very Good				
Good				
Acceptable				204
Un Acceptable	1062	1524	666	

Table 11: Purchase attribute Comparison on TkNN and ImprovedID3withTkNN

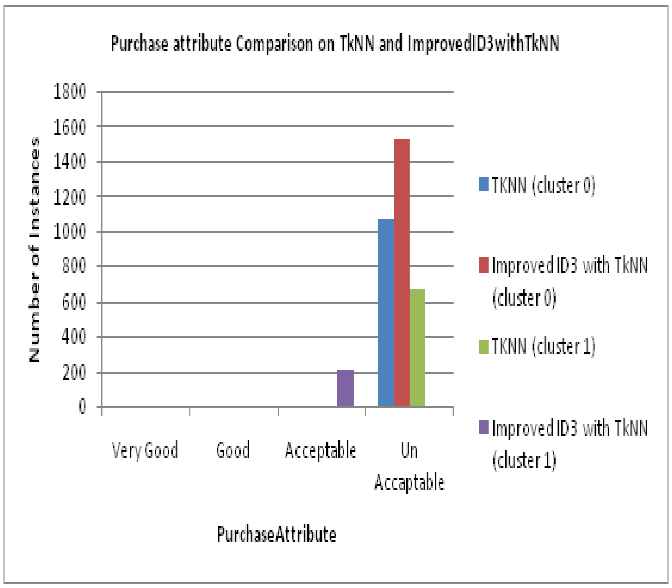

Figure 13: Purchase attribute Comparison on TkNN and ImprovedID3withTkNN

VII. CONCLUSION

This Paper presents ID3, KNN, T*k*NN and our novel improved ID3 with T*k*NN clustering algorithms. We have also executed the same in Weka Tool with Java code and compared the performance of two algorithms based on different Percentage Splits to help the car seller/manufacturer for analyzing their customer views in purchasing a car.

We analyzed the graphical performance analysis between KNN and our novel improved ID3 with T*k*NN clustering algorithms with Classes to Clusters evaluation purchase, safety, luggage booting, persons (seating capacity), doors, maintenance and buying attributes of customer's requirements for unacceptable/acceptable/good/very good ratings of a car to purchase.

REFERENCES

[1] A. Feelders, H. Daniels, M. Holsheimer, "Methodological and Practical Aspects of Data Mining", Information & Management, 271-281, 2000.

[2] Aggarval, C. C., & Yu, P. S,"Finding localized associations in market basket data", IEEE Transactions on Knowledge and Data Engineering, 14, 51–62, 2002.

[3] Alex Berson, Kurt Thearling, Stephen J.Smith, "Building Data Mining Applications for CRM" , kindle edition,eBook, 488 pages,2002.

[4] B. Sun and Morwitz, V.G,"Stated intentions and purchase behavior: A unified mode", International Journal of Research in Marketing,Volume 27(4), 356-366,2010.

[5] B. V. Dasarathy., "Nearest neighbor (nn) norms: Nn pattern classi_cation tech-niques", IEEE Computer Society Press, 1991.

[6] C-L. Huang, M-C. Chen and, C-J. Wang, "Credit scoring with a data mining approach based on support vector machines", Expert System with Applications, 37, 847-856, 2007.

[7] Cabibbo and R. Torlone, "An architecture for data warehousing supporting data independence and interoperability: an architecture for data warehousing", International Journal of Cooperative Information Systems, vol. 10, no. 3, 2001.

[8] Calvanese, G. D. Giacomo, M. Lenzerini, D. Nardi, and R. Rosati, "Data integration in data warehousing", International Journal Of Cooperative Information Systems, vol. 10, no. 3, pp. 237, 2001.

[9] Christoph F. Eick, Nidal Zeidat, and Zhenghong Zhao, "Supervised Clustering – Algorithms and Benefits", 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004), 2004.

[10] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze, "Introduction to information retrieval", Cambridge University Press, 2009.

[11] D. Olson and S. Yong, "Introduction to Business Data Mining". McGraw Hill International Edition, 2006.

[12] Data mining for Business Intelligence. [www.dataminingbook.com]

[13] David J. Hand, Heikki Mannila and Padhraic Smyth, "Principles of Data Mining", MIT Press,Fall 2000.

[14] Douglas, S.Agarwal, D., & Alonso, T,"Mining customer care dialogs for ''daily news". IEEE Transactions on Speech and Audio Processing, 13, 652–660, 2005.

[15] Dunham, M.H.," Data Mining: Introductory and Advanced Topics", Pearson Education Inc.2003.

[16] Fayyad U, "From Data Mining to Knowledge Discovery: An overview", In Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.1996.

[17] Fong, Q.Li, and S. Huang, "Universal data warehousing based on a meta-data modeling approach", International Journal Of Cooperative Information Systems, vol. 12, no. 3, pp. 325, 2003.

[18] Ha, S.H,"Helping online customers decide through web personalization". IEEE Intelligent Systems, 17, 34–43.2002.

[19] IDC & Cap Gemini. "Four elements of customer relationship management". Cap Gemini White Paper.ISBN: 978-0-387-79419-8, 2009.

[20] James A.O. Brien, "Management Information System", Tata Mc Graw Hill publication company Limited, New Delhi, 2009.

[21] Jiang, T., & Tuzhilin, A," Segmenting customers from population to individuals: Does 1-to-1 keep your customers forever", IEEE Transactions on Knowledge and Data Engineering, 18, 1297–1311, 2006.

[22] K. Hian and K.L. Chan, "Going concern prediction using data mining techniques", Managerial Auditing Journal, Vol 19, No 3, 462-476, 2004.

[23] Kalton, K. Wagstaff, and J. Yoo, "Generalized Clustering,Supervised Learning, and Data Assignment," Proceedings of the Seventh International Conference on Knowledge Discovery and DataMining, ACM Press, 2001.

[24] Kantardzic, M.," Data Mining: Concepts, Models, Methods and Algorithms", Wiley-IEEE Press, 2011.

[25] Kerdprasop, and K. Kerdpraso, "Moving data mining tools toward a business intelligence system", Enformatika, vol. 19, pp. 117-122, 2007.

[26] Kilian Q.Weinberger, Lawrence K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification", Journal of Machine Learning Research, 207-244, 2009.

[27] Kubat, M., Hafez, A., Raghavan, V. V., Lekkala, J. R., & Chen, W. K," Item set trees for targeted association querying", IEEE Transaction on Knowledge and Data Engineering, 15, 1522–1534, 2003.

[28] Lapersonne, G. Laurent and J-J Le Goff, "Consideration sets of size one: An empirical investigation of automobile purchases", International Journal of Research in Marketing 12, 55-66, 1995.

[29] Leontin, T. L., Moldovan, D., Rusu, M., Secară, D., Trifu, C,"Data mining on the real estate market", Revista Informatica Economică, nr. 4 (36), 2005.

[30] Liqiang and J. Howard, "Interestingness measures for data mining: a survey", ACM Computing Surveys, vol. 38, no. 3, pp. 1-32, 2006.

[31] M. Panda and M. Patra,"A novel classification via clustering method for anomaly based network intrusion detection system", International Journal of Recent Trends in Engineering, 2:1–6, 2009.

[32] M.R.Lad, R.G.Mehta, D.P.Rana, "A Noval Tree Based Classification", [IJESAT] International Journal of Engineering and Advanced Technology Volume-2, Issue-3, 581 – 586 may 2012.

[33] Moutinho, L., Davies, F. and Curry, B,"The impact of gender on car buyer satisfaction and loyalty". Journal of Retailing and Consumer Services 3(3), 135-144, 1996.

[34] M. Matteucci,"A Tutorial on Clustering Algorithms", 2008. [http://home.dei.polimi.it/matteucc/Clustering/tutorial_html].

[35] Ming, H., Wenying, N. and Xu, L, "An improved decision tree classification algorithm based on ID3 and the application in score analysis", Chinese Control and Decision Conference (CCDC), pp1876-1879, 2009.

[36] Mitra, S., Pal, S. K., & Mitra, P,"Data mining in soft computing framework: A survey". IEEE Transactions on Neural Networks, 13, 3–14, 2002.

[37] R. Krakovsky and R. Forgac,"Neural network approach to multidimensional data classification via clustering", Intelligent Systems and Informatics (SISY), 2011 IEEE 9th International Symposium on, 169–174, IEEE2011.

[38] Rahul A. Patil, Prashant G. Ahire, Pramod. D. Patil, Avinash L. Golande ,"A Modified Approach to Construct Decision Tree in Data Mining Classification" , International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 1, July 2012.

[39] Rosset, S., Neumann, E., Eick, U., & Vatnik, N,"Customer lifetime value models for decision support", Data Mining and Knowledge Discovery, 7, 321–339, 2003.

[40] R. Nayak and T. Qiu, "A data mining application: analysis of problems occurring during a software project development process", International Journal Of Software Engineering & Knowledge Engineering, vol.15, no. 4, pp. 647-663, 2005.

[41] S. Bongsik, "An exploratory investigation of system success factors in data warehousing", Journal of the Association for Information Systems, vol. 4, pp. 141-168, 2003.

[42] S. Lee, S. Hong, and P. Katerattanakul, "Impact of data warehousing on organizational performance of retailing firms", International Journal of Information Technology & Decision Making, vol. 3, no. 1, pp. 61-79, 2004.

[43] Sen and A. P. Sinha, "A comparison of data warehousing methodologies", Communications of The ACM, vol. 48 Issue 3, pp. 79-84, 2005.

[44] Sun and Morwitz, V.G., Stated intentions and purchase behavior: A unified model. International Journal of Research in Marketing, Volume 27(4), 356-366, 2010.

[45] Su, C. T., Hsu, H. H., & Tsai, C. H,"Knowledge mining from trained neural networks", Journal of Computer Information Systems, 42, 61–70,2002.

[46] T. Hertz, A. Hillel, and D. Weinshall, "Learning a Kernel Function for Classification with Small Training Samples," Proc. ACM Int'l Conf. Machine Learning, 2006.

[47] Transportation and Economic Development. [https://people.hofstra.edu/geotrans/eng/ch7en/conc7en/ch7c1en.html]

[48] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases", Artificial Intelligence. AI Magazine, pp. 37-54, 1996.

[49] UCI Machine Learning Repository – [http://mlearn.ics.uci.edu/databases]

[50] W. Smith, "Applying data mining to scheduling courses at a university", Communications Of AIs; vol. 2005, no. 16, pp. 463-474, 2005.

[51] W. Hugh, A. Thilini, Jr. Matyska, and J. Robert, "Data warehousing stages of growth", Information Systems Management; vol. 18, no. 3, pp.42-51, 2001.

[52] Wai-Ho Au, Member, IEEE, Keith C. C. Chan, Andrew K.C. Wong, Fellow, IEEE,and Yang Wang, Member, IEEE ,"Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data", Sep. 15, 2004.

[53] W. Smith, "Applying data mining to scheduling courses at a university", Communications Of AIs; vol. 2005, no. 16, pp. 463-474, 2005.

[54] WEKA Software, The University of Waikato. [http://www.cs.waikato.ac.nz/ml/weka].

BIOGRAPHY

M.Jayakameswaraiah,Received his Master of Computer Applications Degree from Sri Venkateswara University,Tirupati, Andhrapradesh, INDIA in 2009 and Currently Pursuing Ph.D in Computer Science from Sri Venkateswara University ,Tirupati, Andhrapradesh, INDIA.The Research fields of interest are Data Mining, Cloud Computing, Software Engineering and Data Base Management System etc. E-Mail: mjayakameswaraiah@gmail.com

Prof.S.Ramakrishna, Professor in Department of Computer Science, Sri Venkateswara University,Tirupati, Andhrapradesh,India.His Areas of interest include Data Mining, Software Engineering,Data Base Management System, Image Processing and Machine learning etc. E-Mail: drsramakrishna@yahoo.com