International Journal of Engineering Sciences Paradigms and Researches (IJESPR)
(Vol. 13, Issue 01) and (Publishing Month: May 2014)
(An Indexed, Referred and Impact Factor Journal)
ISSN (Online): 2319-6564
www.ijesonline.com

Clustering Based Classification in E-Commerce

Ms Goldy Rana[1] and Mrs Silky Azad(Guide)[2]

[1]M.Tech. Student and [2]Assistant Professor
[1,2]Computer Science and Engineering Department
[1,2]Samalkha Group of Institutions, Hathwala, Panipat (Haryana)
[1]goldyrana1991@gmail.com
[2]silkyzd15@gmail.com

Abstract

Electronic commerce offers a huge variety of application areas for Artificial Intelligence. A major requirement addressed by current research is the availability of competent virtual sales agents that guide the customers through the vast space of available products, services, and other opportunities. One of the most important problems today is to achieve an appropriate communication between the customer and the virtual sales agent. Such a communication should be similar to a communication with a real sales agent in a bricks and-mortal store: the agent must ask appropriate questions concerning the customer's product requirements and should at a certain point provide appropriate product information. Clustering is to find clusters from a large amount of data samples, maximizing the similarity of intra-cluster samples and minimizing that between inter-cluster samples. That is to say, clustering is to discover densely populated regions in the whole data space, and every dense region is a cluster. A density functions and calculates the density of every data sample. Then dense regions, which are just clusters, can be found according to the densities of data samples. In his paper, an algorithm is design that will make clusters of same product from massive dataset of products on the basis of property of product.

Keywords: *E-Commerce, Data- Mining, Clustering, REP Tree, K-Mean.*

I. Introduction

Data mining is concerned with the discovery and extraction of latent knowledge from a database. Typically, this knowledge is classified into rules and patterns that can help an analyst in analysis and decision making processes. Data mining has been used for a wide range of applications ranging from decision support systems in business applications to analysis tools in scientific applications. The purpose of data mining can be predictory (decision support), generative (create new/improved designs), or explanatory (scientific analysis) [1].

Web (usage) mining is the analysis of (user behaviour) data in Web-based systems. The database is the access log created by a Web server. The fact that only activities are recorded makes Web usage mining different from data mining in general. Each Web request of any text document or other type of resource is recorded in the access log. Web log entries reflect activities – the requestor, access type, access time, and requested resource are recorded. In education-specific terms the learner, form of activity/interaction, access time, and content item are recorded [1].

A range of classical data mining techniques exist that support the extraction of rules and patterns from a database [2] [3]:

- Usage statistics are usually not considered as data mining techniques. However, they often form the starting point of evaluations. For Web-based systems, usage is captured in simple statistical measures such as total number of visits, number of visits per page, and so forth. Tracking features of most e-learning platforms are based on these measures.

- Classification and prediction are related techniques. Classification predicts class labels, whereas prediction predicts continuous-valued functions. A model is used to analyse a sample. The result of this learning step is then applied. Regression is a typical form of prediction.

- Clustering groups mutually similar data items. In contrast to classification, the class labels are not pre-given. The learning process is called unsupervised in this technique. Pattern recognition is a typical example.

- Association rules are interesting relationships discovered among the set of data items. A typical example is purchasing analysis, which can identify item pairs frequently purchased together.

- Sequential pattern analysis is applied if events are captured in a database over a period of time. Frequently occurring patterns are extracted. Web usage or sales transaction patterns are typical examples.

- Time series, the analysis of the variance of patterns and rules over time, are important since they allow the analyst to evaluate changing and varying behaviour. Often, a session, which is a period of uninterrupted usage, is the basic unit of analysis.

The understanding of levels of abstraction of knowledge and of languages to express knowledge is critical for the success of the data mining technology. Concepts, e.g. learning activities and interactions, have to be clarified.

Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering [4]. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another and very dissimilar to object in other clusters. Dissimilarity is due to the attributes values that describe the objects [5]. The objects are grouped on the basis of the principle of optimizing the intra-class similarity and reducing the inter-class similarity to the minimum. First of all the set of data is portioned into groups on the basis of data similarity (e g by clustering) and the then assigning labels to the comparatively smaller number of groups [5].

II. E-Commerce and Data Mining

The advent of e-commerce revolutionized every industry. Every aspect of commerce, from sales pitch to final delivery, could be automated and made available 24 hours a day, all over the world. B2B solutions carried this one step further, allowing vertical partnerships and co-branding. Businesses found a new incentive to bring their data into the digital age [6]. And dynamic content allowed the first truly personalized, interactive websites to come into being, all through the magic of e-commerce.

Data are collected electronically, rather than manually, so less noise is introduced from manual processing. E-commerce data are rich, containing information on prior purchase activity and detailed demographic data. In addition, some data that previously were very difficult to collect now are accessible easily. For example, e-commerce systems can record the actions of customers in the virtual "store," including what they look at, what they put into their shopping cart and do not buy, and so on. Previously, in order to obtain such data companies had to trail customers (in person), surreptitiously recording their activities, or had to undertake complicated analyses of in-store videos. It was not cost-effective to collect such data in bulk, and correlating them with individual customers is practically impossible. For e-commerce systems massive Amounts of data can be collected inexpensively [6].

III. Clustering Techniques

a. REP Tree

Basically Reduced Error Pruning Tree ("REPT") is a fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. The basic of pruning of this algorithm is it used REP with back over fitting. It kindly sorts values for numerical attribute once and it handling the missing values with embedded method by C4.5 in fractional instances. In this algorithm we can see it used the method from C4.5 and the basic REP also count in it process [7].

REP Tree algorithm [8] is based on the principle of calculating the information gain with entropy and reducing the error arising from variance [9]. This method is firstly suggested by Quinlan [10]. With the help of this method, complexity of decision tree model is decreased by "reduced error pruning method" and the error arising from variance is reduced [8].

Decision tree is a tree formed data structure that verifies divide and rule approach. Decision tree is used for supervised learning. It is a tree structured model in which the local region is found recursively, with a set of division in a few steps. Decision tree consists of inner decision node and outer leaf. Every decision node *m verifies* an $f_m(x)$ test function whose discrete value is related to branches. Test function is performed in each node for an input and one of the branches is selected according to the result. This process starts in root and continues recursively until a leaf node is reached; the value written on the leaf produces the output [9, 11].

Decision trees are one of the most widely used classifiers on classifying problems. It is easily understood and configured compared to other methods [8,9,11]. Let Y and X be the discrete variables that have the values $\{y_1, \ldots, y_n\}$ ve $\{x_1, \ldots, x_n\}$. In this case, entropy and conditional entropy of Y are calculated as shown in equation (1) and (2). After that, information gain of X is calculated as shown in equation (3).

$$H(Y) = -\sum_{i=1}^{k} P(y = y_i) \log P(Y = y_i) \quad (1)$$

$$H(Y|X) = -\sum_{i=1}^{l} P(X = x_i) H(Y|X = x_i) \quad (2)$$

$$IG(Y;X) = H(Y) - H(Y|X) \quad (3)$$

In decision trees, pruning is done in two ways. These are pre-pruning and post-pruning. If the number of instances that reach a node is lowers than the percentage of the training set, that node is not divided. It is considered that varience of the model which is generated by the training with a small number of instances and accordingly the generalization error will increase. For this reason, if the expansion of the tree is stopped when building the tree, then this is called pre-pruning.

Another way of building simple tress is post-pruning. Generally, post-pruning gives better results than pre-pruning in practice [11]. Since the tree does not take steps backward and continues to expand steadily while it is being built, the variance increases. Post-pruning is a way to avoid this situation. In order to do this, firstly, unnecessary sub-trees should be found and pruned.

In post-pruning, the tree is expanded until all the leaves are pure and there is no error in training set. After that, we find the sub-trees that lead to memorizing and prune them. In order to this, we firstly use a major part of training set as growing set and the remaining part as pruning set. Later, we replace each sub-tree with a leaf that is trained by the instances which are covered by the training set of that sub-tree and then we compare these two options on pruning set. If the leaf does not lead to more errors on pruning set, we prune the sub-tree and use the leaf, otherwise we keep the sub-tree [12]. When we compare and contrast pre-pruning and post-pruning, we see that pre-pruning produces faster trees, on the other hand, post-pruning produces more successful trees [8].

b. K-Mean Clustering

K-means clustering is a partitioning based clustering technique of classifying/grouping items into k groups (where k is user specified number of clusters). The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid. A centroid (also called mean vector) is "the center of mass of a geometric object of uniform density". Although K-means is simple and can be used for a wide variety of data types, it is quite sensitive to initial positions of cluster centers [13]. There are two simple approaches to cluster center initialization i.e. either to select the initial values randomly, or to choose the first k samples of the data points. As an alternative, different sets of initial values are chosen (out of the data points) and the set, which is closest to optimal, is chosen. However, testing different initial sets is considered impracticable criteria, especially for large number of clusters, Ismail et al (1989). Therefore, different methods have been proposed in literature by Pena et al. (1999). Also, the computational complexity of original K-means algorithm is very high, especially for large data sets. Computer science has been widely adopted in different fields like agriculture. One reason is that an enormous amount of data has to be gathered and analyzed which is very hard or even impossible without making use of computer systems [13].

In data mining, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results in a partitioning of the data space into Voronoi cells. Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets (k ≤ n) S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares (WCSS) [14]:

$$\arg min \sum_{i=1}^{k} \sum_{x_{j \in s_i}} \|x_j - \mu_i\|^2$$

where μi is the mean of points in Si.

The basic step of k-means clustering is simple. In the beginning, we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids. Then the K means algorithm will do the three steps below until convergence. Iterate until stable (= no object move group) [15]:

1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids .
3. Group the object based on minimum distance (find the closest centroid)

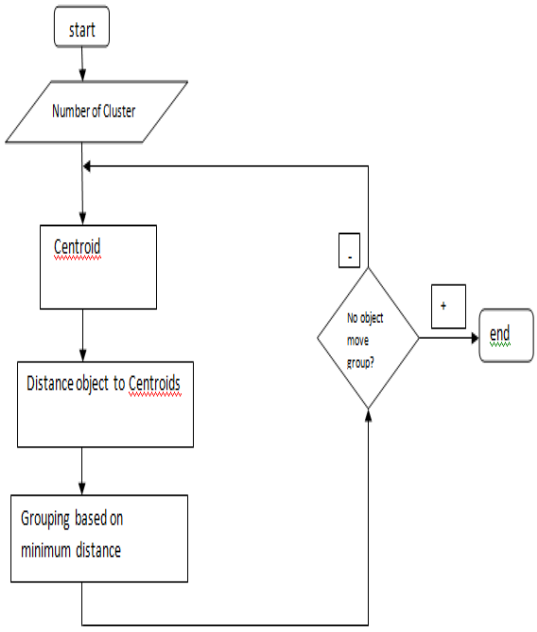

Figure 1: K-Means Clustering Process [15].

IV. Proposed Work

In this work an algorithm is proposed that group the items on the basis of their attributes then classify the clusters. In other words, the proposed algorithms firstly cluster the items on the basis of property i.e. attributes available for the dataset. The clustering is performed is the K-Mean clustering. Then this clustered data is classified using the REP tree which is already explained in section III. It means the proposed algorithm is the hybrid algorithm of K-MEAN clustering and the REP tree classification. It can be explained by the following algorithm.

a. **Proposed Algorithm**

1. Define K i.e. number of clusters.
2. Select K data elements as centroid randomly.
3. For each element in data set
4. For k=1:K
5. Calculate the distance with k centroid say d(i,k).
6. End for
7. End for
8. Create the clusters on the basis of minimum distance i.e. Min(d(i,:)).
9. Calculate the mean and select element as new centroid in each group
10. If new centroid != existing centroid

11. Go to step 3
12. End if
13. For j=1:K
14. Calculate the information gain with entropy in the cluster j.

$$H(Y) = -\sum_{i=1}^{k} P(y = y_i) \log P(Y = y_i) \quad (1)$$

$$H(Y|X) = -\sum_{i=1}^{l} P(X = x_i) H(Y|X = x_i) \quad (2)$$

$$IG(Y; X) = H(Y) - H(Y|X) \quad (3)$$

15. If variance of element is not less than the training variance
16. Divide the elements of cluster
17. Go to step 13
18. End if
19. End for

This algorithm can also be explained by the following flowchart.

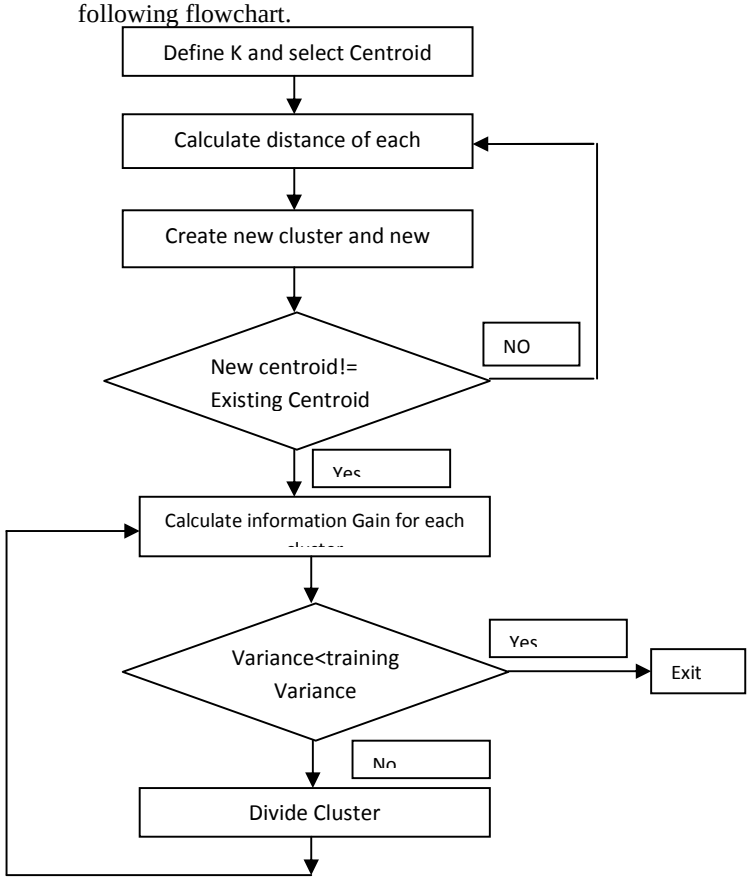

Figure 2: Proposed Algorithm using Flowchart

V. Results

The dataset used to analyze the proposed algorithm over WEKA is the clothing dataset. This dataset is downloaded from the "https://github.com/vincentarelbundock/Rdatasets/blob/master/csv/Ecdat/Clothing.csv". The data set is in CSV i.e. comma separated value format. This data set contains 400 instances each having 14 attributes. The attributes are as: srno, tsales, sales, margin, nown, nfull, npart, naux, hoursw, hourspw, inv1, inv2, ssize, start. These attributes within the dataset explains the size and the sale with corresponding profit of the t-shirt sale. This is basically data set that identifies the sales of a product along with other characteristics.

a. Performance Evaluation Parameters

1. Root-Mean-Square

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. Basically, the RMSD represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSD is a good measure of accuracy, but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale-dependent [16]; the general equation for root mean square error (RMSE) is [17]:

$$RMSE = \sqrt{\frac{(\sum_{i=1}^{n}(a_i - b_i)^2)}{n}}$$

2. Relative Absolute Error

The **relative absolute error** is very similar to the relative squared error in the sense that it is also relative to a simple predictor, which is just the average of the actual values. In this case, though, the error is just the total absolute error instead of the total

squared error. Thus, the relative absolute error takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor. Mathematically, the **relative absolute error** E_i of an individual program *i* is evaluated by the equation [18]:

$$E_i = \frac{\sum_{j=1}^{n} |P_{(ij)} - T_j|}{\sum_{j=1}^{n} |T_j - \bar{T}|}$$

Where $P_{(ij)}$ is the value predicted by the individual program *i* for sample case *j* (out of *n* sample cases); T_j is the target value for sample case *j*; and $\bar{T}$ is given by the formula:

$$\bar{T} = \frac{1}{n} \sum_{j=1}^{n} T_j$$

For a perfect fit, the numerator is equal to 0 and E_i = 0. So, the E_i index ranges from 0 to infinity, with 0 corresponding to the ideal.

3. Relative Square Error

The root relative squared error is relative to what it would have been if a simple predictor had been used. More specifically, this simple predictor is just the average of the actual values. Thus, the relative squared error takes the total squared error and normalizes it by dividing by the total squared error of the simple predictor. By taking the square root of the relative squared error one reduces the error to the same dimensions as the quantity being predicted. Mathematically, the root relative squared error E_i of an individual program *i* is evaluated by the equation [19]:

$$\boldsymbol{E_i = \frac{\sum_{j=1}^{n} |P_{(ij)} - T_j|}{\sum_{j=1}^{n} |T_j - \bar{T}|}}$$

Where $P_{(ij)}$ is the value predicted by the individual program *i* for sample case *j* (out of *n* sample cases); T_j is the target value for sample case *j*; and $\bar{T}$ is given by the formula [19]:

$$\bar{T} = \frac{1}{n} \sum_{j=1}^{n} T_j$$

For a perfect fit, the numerator is equal to 0 and E_i = 0. So, the E_i index ranges from 0 to infinity, with 0 corresponding to the ideal.

4. Mean Absolute Error

In statistics, the **mean absolute error (MAE)** is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by [20]:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |f_i - y_i| = \frac{1}{n} \sum_{i=1}^{n} |e_i|$$

As the name suggests, the mean absolute error is an average of the absolute errors $e_i = |f_i - y_i|$, where f_i is the prediction and y_i the true value. Note that alternative formulations may include relative frequencies as weight factors [20].

The mean absolute error is a common measure of forecast error in time series analysis, where the terms "mean absolute deviation" is sometimes used in confusion with the more standard definition of mean absolute deviation. The same confusion exists more generally.

5. Correlation Coefficient

The correlation coefficient of two variables in a data sample is their covariance divided by the product of their individual standard deviations. It is a normalized measurement of how the two are linearly related [21]. Formally, the sample correlation coefficient is defined by the following formula, where s_x and s_y are the sample standard deviations, and s_{xy} is the sample covariance [21].

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

Similarly, the population correlation coefficient is defined as follows, where σ_x and σ_y are the population standard deviations, and σ_{xy} is the population covariance.

$$p_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

If the correlation coefficient is close to 1, it would indicates that the variables are positively linearly related and the scatter plot falls almost along a

straight line with positive slope. For -1, it indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope. And for zero, it would indicates a weak linear relationship between the variables [21].

The proposed and REP Tree algorithms are simulated over the WEKA tool using the clothing dataset. Their performance analysis is shown in the following table.

Table 1 Parameter Analysis

Parameter Name	Proposed	REP Tree
Mean absolute error	0.3289	2.2543
Root mean squared error	2.6562	5.8994
Relative absolute error	3.9739%	27.2344%
Root relative squared error	19.9394%	44.2859%

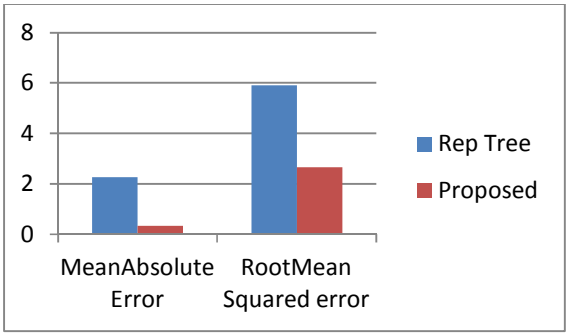

Figure 3 Comparisons of MAE and RMSE

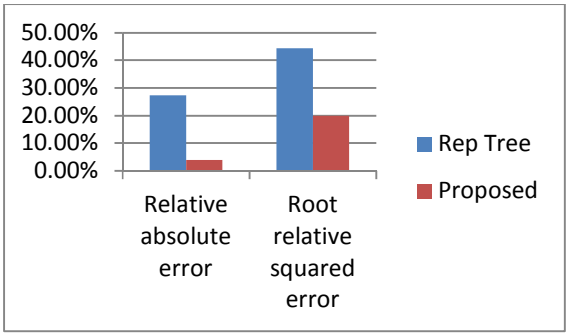

Figure4 Comparisons of RAE and RRSE

The figure 3 shows that the proposed algorithm decreases the mean absolute error as well as the root mean square error. The relative absolute error and the root relative absolute error are also decreased as shown in the figure 4. The decrease in error results in accurate classification. So the proposed algorithm cluster the items ant classifies them on the basis of their attributes more accurately. The relation between the items is derived from the attributes present in the dataset. This relation is used to cluster and classify the items. So the proposed algorithm fully depends on the attributes present in the dataset.

VI. Conclusion

The paper proposes an algorithm that groups the items on the basis of their attributes then classify the clusters. In other words, the proposed algorithms firstly cluster the items on the basis of property i.e. attributes available for the dataset. The clustering is performed is the K-Mean clustering. Then this clustered data is classified using the REP tree. In other words the proposed algorithm is the hybrid algorithm of K-MEAN clustering and the REP tree classification. The proposed algorithm is compared with the REP Tree algorithm using the WEKA tool. The comparison is done over clothing dataset downloaded from internet. The proposed algorithm decreases the mean absolute error as well as the root mean square error. The relative absolute error and the root relative absolute error are also decreased. The decrease in error results in accurate classification. So the proposed algorithm cluster the items ant classifies them on the basis of their attributes more accurately. In future the algorithm can also be analyzed over other e-commerce datasets. The algorithm can be modified to be applicable for other fields like medical. The neural network can also be added to the proposed algorithm to enhance the performance.

References

[1] Pahl, C. (2004). Data mining technology for the evaluation of learning content interaction. International Journal on E-learning, 3(4), 47.

[2] Agrawal, R., & Srikant, R. (1995, March). Mining sequential patterns. In Data Engineering, 1995. Proceedings of the Eleventh International Conference on(pp. 3-14). IEEE.

[3] Cooley, R., Tan, P. N., & Srivastava, J. (2000). Discovery of interesting usage patterns from web

data. In Web Usage Analysis and User Profiling (pp. 163-182). Springer Berlin Heidelberg.

[4] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya, "Comparison the various clustering algorithms of weka tools", International Journal of Emerging Technology and Advanced Engineering ,Volume 2, Issue 5, May 2012.

[5] Vishal Shrivastava, Prem narayan Arya, (October 2012) A Study of Various Clustering Algorithms on Retail Sales Data, International Journal of Computing, Communications and Networking, Volume 1, No.2.

[6] Sanchati, R., Patidar, P. C., & Kulkarni, G.(2011) Path Breaking Case Studies in E-commerce using Data Mining. International Journal of Computer Technology and Electronics Engineering, 1.

[7] Mohamed, W., Salleh, M. N. M., & Omar, A. H. (2012, November). A comparative study of Reduced Error Pruning method in decision tree algorithms. In Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on (pp. 392-397). IEEE.

[8] Zontul, M., Dogan, G., Aydin, F., Sener, S., & Kaynar, O. (2013). Wind Speed Forecasting Using Reptree And Bagging Methods In Kirklareli-Turkey. Journal of Theoretical & Applied Information Technology, 56(1).

[9] Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques – 2nd ed.. the United States of America, Morgan Kaufmann series in data management systems.

[10] Quinlan, J. R. (1987). Simplifying decision trees. International journal of man-machine studies, 27(3), 221-234.

[11] Alpaydın E (2004) Introduction to Machine Learning. The MIT Press, Printed and bound in the United States of America. ISBN: 0-262-01211-1.

[12] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast Discovery of Association Rules. Advances in knowledge discovery and data mining, 12(1), 307-328.

[13] Ritu Sharma(Sachdeva) (2014) K-Means Clustering in Spatial Data Mining using Weka Interface ,International Conference on Advances in Communication and Computing Technologies (ICACACT) 2012 Proceedings published by International Journal of Computer Applications® (IJCA).

[14] Namita Bhan (Comparative Study Of EM And K-Means Clustering Techniques In WEKA Inter-Face, International Journal of Advanced Technology & Engineering Research (IJATER), Volume 3, Issue 4, July 2013.

[15] Sapna Jain, (2010) K-MEANS Clustering Using WEKA Interface, Proceedings of the 4th National Conference; INDIACom-2010 Computing For Nation Development, February 25 – 26.

[16] http://en.wikipedia.org/wiki/Root-mean-square_deviation.

[17] http://www.dtic.mil/dtic/tr/fulltext/u2/a302958.pdf.

[18] http://www.gepsoft.com/gxpt4kb/Chapter10/Section2/SS15.htm.

[19] http://www.gepsoft.com/gxpt4kb/Chapter10/Section1/SS07.htm.

[20] http://en.wikipedia.org/wiki/Mean_absolute_error.

[21] http://www.r-tutor.com/elementary-statistics/numerical-measures/correlation-coefficient.