ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 3, pp: (249-254), Month: July - September 2015, Available at: www.researchpublish.com

Clustering of Categorized Text Data Using Cobweb Algorithm

[1]Kavita, [2]Pallavi Bedi

[1]Research Scholar, [2]Assistant professor CSE Deparment in Prannath Parnami Institute of Hisar, India

Abstract: **The objective of clustering is to partition an unstructured set of objects into clusters (groups). Initially the data is not structured. In clustering distance & the similarity between the objects is consider. One often wants to group similar objects in same cluster and dissimilar in different clusters. Clustering is a widely studied data mining problem in text domain. In this paper we make use of a database 'Labor Dataset' in arff (attribute relation file format) containing 17 attributes and 57 instances to perform an clustering and classification techniques of data mining. We get results with simple classification technique (using naïve bayes classifier) and clustering technique (using cobweb algorithm), based upon various parameters using WEKA (Waikato Environment for Knowledge Analysis), a Data Mining tool. The results of the experiment show that clustering and classification gives promising results with utmost accuracy rate and robustness even when the data set is containing missing values.**

Keywords: **Data Mining; Naïve bayes classifier; Cobweb algorithm; WEKA; labor dataset.**

1. INTRODUCTION

Data mining is the process of extraction of useful information from the large amount of the data. Extracting the new information which is unknown for us. In data mining the initial data is noisy & inconsistent in nature ie the data is redundant & contain missing values. So in data mining firstly done the data cleaning then integration ie collection of all the data , data selection, data transformation, data preprocessing.[1] Data mining is used in banks, institutes, companies, marketing, government agencies , universities etc because these organizations can contain large amount of the data. It is known as knowledge discovery of databases ie KDD. The techniques clustering, classification, preprocessing, association rule mining, visualization are done processing in data mining. Clustering is one of the most important data mining or text mining algorithm that is used to group similar objects together. In clustering consider the data set for eg labor, weather, iris etc in arff format. The data is contain the attributes & instances. So the objects which have same attributes can group in to one cluster because these have similarity & the objects which are not similar are assigned to other cluster. In clustering dividing the text the data in different classes for eg the data of sports class are data related to cricket, data related to basketball etc. So in this manner the data is classified in different categories. Clustering is used in extraction of useful data in data mining, machine learning, reorganization of patterns, image analysis fields [2]. A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. In clustering the clustering method is best which can't left the outliers outside the clusters. The clustering algorithms can assigned all the objects in to their clusters. Outliers are the deviations from normal values. Outliers are the value which is outside of the clusters.

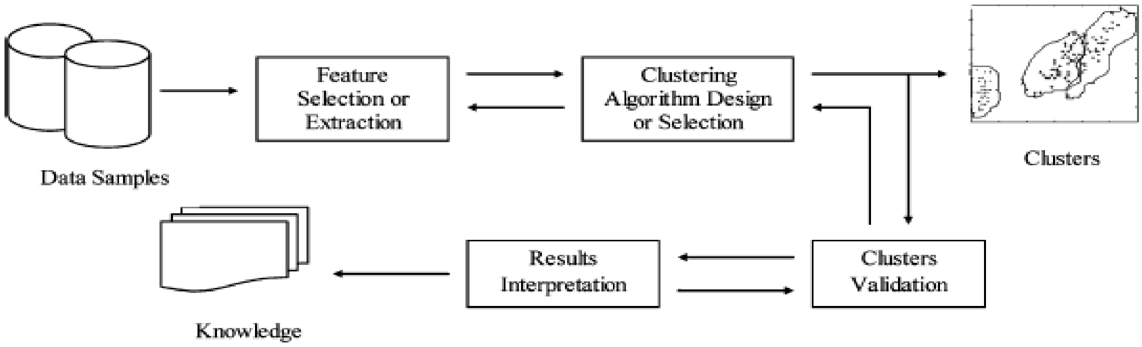

Data Samples
Feature Selection or Extraction
Clustering Algorithm Design or Selection
Clusters
Knowledge
Results Interpretation
Clusters Validation

Fig 1. Clustering diagram

Cobweb algorithm is used for clustering because it is fast & give good results. In cobweb merging & splitting is done on the objects for adding the object in to the tree. In cobweb algorithm the result is in dendrogram i.e tree [3].It is based on category utility function. In cobweb unseen objects is entered in to the tree with the help of merging & splitting techniques with out changing the tree. It is bidirectional because if the tree is merged then previous split is also done. The proposed work will focus on clustering and classification techniques [4]. Classification has been identified as an important problem in the emerging field of data mining. Given our goal of classifying large data sets, we focus mainly on naive bayes classifiers. Naïve bayes classifier can produce better accuracy results with the missing values. In this paper WEKA (Waikato Environment for knowledge analysis) machine learning tool is used for performing clustering and classification algorithms. The dataset used in this paper is labor dataset in arff format containing 17 attributes and 57 instances with two classes good or bad.

2. PROBLEM STATEMENT

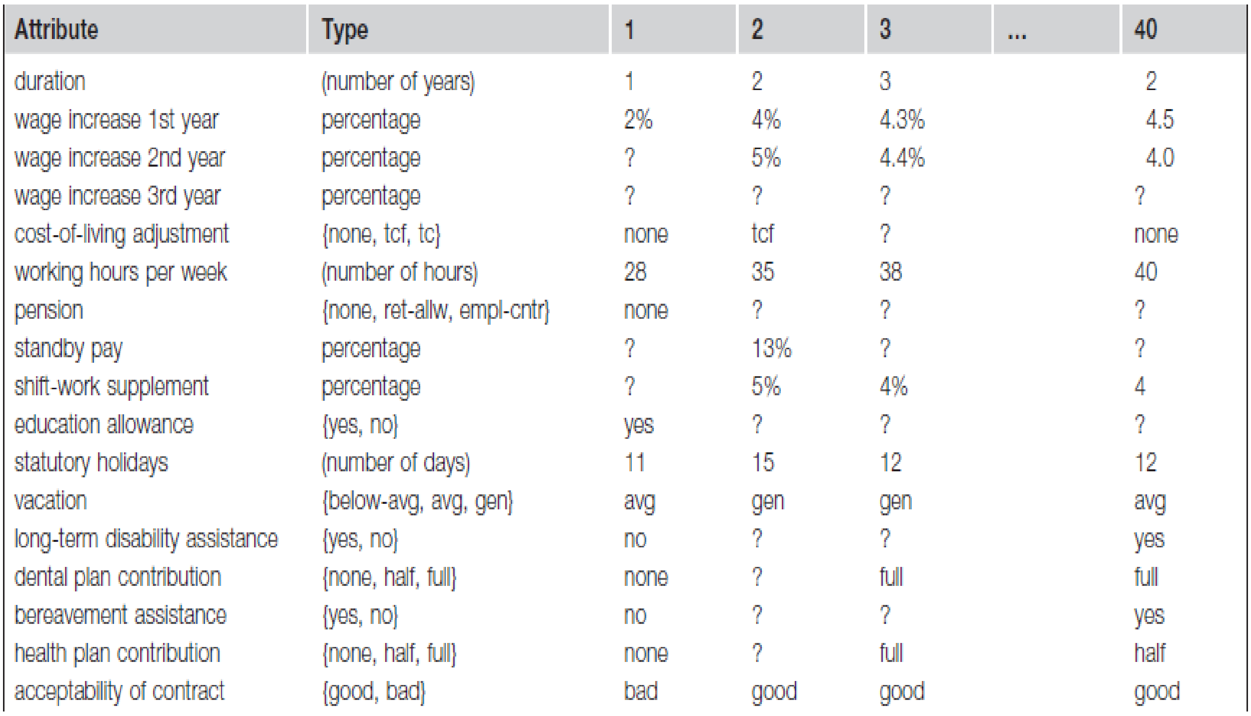
In our problem the labor dataset is used in arff format (attribute relation file format) contains 17 attributes and 57 instances containing missing values or nominal values i.e which is represented in class. For eg in labor dataset the class is represented in good ,bad. And the dataset contain numeric values. Firstly we can done preprocessing using unsupervised preprocessing technique i.e dividing continuous data into discrete form. Done the classification of the labor dataset with the help of naïve bayes classifier and then done clustering with the help of cobweb algorithm. The cobweb algorithm is incremental hierarchical algorithm which is created dendrogram ie tree like structure.We use weka tool for the data preprocessing, data clustering and data classification. Weka is data mining tool. It supports graphical user interface & the Weka implementation is done java language.

3. PROPOSED WORK

A. Dataset used:

We can take the labor negotiation dataset in arff format with 17 attributes and 57 instances. The attributes contain missing values and nominal values that is represented in classes. Missing values are represented by ?. We can classify the labor dataset with NAÏVE BAYES classification algorithm and done clustering with COBWEB algorithm. Before Clustering and classification preprocessing is done on the labor dataset with the help of unsupervised preprocessing using equal frequency method. The attributes of the labor dataset have numeric, nominal values ie class {bad, good}.

Table 1: Labor dataset



Attribute	Type	1	2	3	...	40
duration	(number of years)	1	2	3		2
wage increase 1st year	percentage	2%	4%	4.3%		4.5
wage increase 2nd year	percentage	?	5%	4.4%		4.0
wage increase 3rd year	percentage	?	?	?		?
cost-of-living adjustment	{none, tcf, tc}	none	tcf	?		none
working hours per week	(number of hours)	28	35	38		40
pension	{none, ret-allw, empl-cntr}	none	?	?		?
standby pay	percentage	?	13%	?		?
shift-work supplement	percentage	?	5%	4%		4
education allowance	{yes, no}	yes	?	?		?
statutory holidays	(number of days)	11	15	12		12
vacation	{below-avg, avg, gen}	avg	gen	gen		avg
long-term disability assistance	{yes, no}	no	?	?		yes
dental plan contribution	{none, half, full}	none	?	full		full
bereavement assistance	{yes, no}	no	?	?		yes
health plan contribution	{none, half, full}	none	?	full		half
acceptability of contract	{good, bad}	bad	good	good		good

Table 2. The result of the discretization of attribute wage increase first year shows frequency count & weight.

Sr no.	Wage increase in first year	Frequency Count	weight
1.	(-inf-2.05]	10	10.0
2.	(2.05-2.65]	5	5.0
3.	(2.65-3.25]	6	6.0
4.	(3.25-3.6]	5	5.0
5.	(3.6-4.15)	7	7.0
6.	(4.15-4.4]	1	1.0
7.	(4.4-4.55]	9	9.0
8.	(4.55-5.35]	6	6.0
9.	(5.35-6.2]	4	4.0
10.	(6.2-inf)	3	3.0

B. Algorithms:

1. Naive bayes: Naïve bayes classifier is used when the amount of the input is high means it can work accurately on the large amount of the data. It can give highly accurate results with large data.In naïve bayes classifier the record can contain the attributes.[5] So consider each attribute in each class separately & then make the training easy & fast in speed. The naïve bayes algorithm is used to filter spam mails. The spam mail contain virus in the internet.

2. Cobweb cluster algorithm**:** The cobweb clustering algorithm is incremental hierarchical algorithm which can incrementally add the elements & it is based on the category utility function**.** The cobweb algorithm can large number of the elements in the tree so we can have a cutoff feature to cut the no. of elements from the tree like structure. So it can reduce the tree size. In the cobweb method the splitting & merging of the elements is done. So it can done bidirectional search. In cobweb we can split the tree merged tree to form multiple clusters. Where as In the k means the cluster can added the element which is nearest to the cluster centre.

$$CU(C_1, C_2, \ldots, C_k) = \frac{\sum_l \Pr[C_l] \sum_i \sum_j \overbrace{\left(\Pr[a_i = v_{ij} \mid C_l]^2 - \Pr[a_i = v_{ij}]^2\right)}^{\substack{\text{Improvement in probability estimate} \\ \text{because of instance cluster assigment}}}}{k}$$

If each instance in its own cluster:

$$\Pr[a_i = v_{ij} \mid C_l] = \begin{cases} 1 & v_{ij} = \text{actual value of instance} \\ 0 & \qquad \text{otherwise} \end{cases}$$

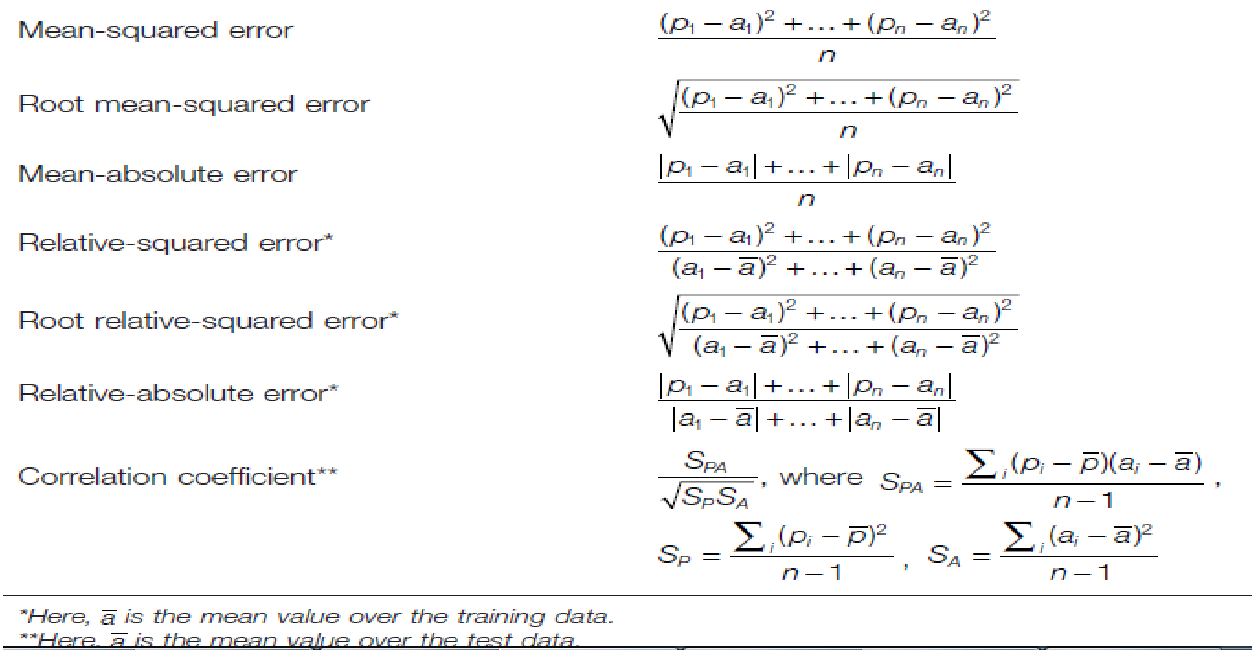
C. Measures for performance evaluation: To evaluate the performance sensitivity and specificity are consider. The sensitivity is true positive rate it is determined by dividing the true positives by the total no. of positives ie TP+FN. The specificity is determined by dividing the true negatives by total no. of positives ie .TN+FP i.e true negatives and false positives.

$$\textit{Sensitivity or recall } TPR = \frac{TP}{TP+FN}$$

$$\textit{specificit y} = \frac{TN}{FP+TN}$$

Naïve bayes terms : These terms i.e kappa statistic , mean squared error, root mean squared error , mean absolute error , relative squared error , relative absolute error are used in the results of naïve bayes algorithm. In this table p1,p2,p3…pn are the estimated values which are estimated from real values & the a1,a2,a3…. an are real values or accurate values.

Table 3. Performance measures for naïve bayes terms



Mean-squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$
Root mean-squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$
Mean-absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{n}$
Relative-squared error*	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}$
Root relative-squared error*	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}}$
Relative-absolute error*	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{\lvert a_1-\bar{a}\rvert+\ldots+\lvert a_n-\bar{a}\rvert}$
Correlation coefficient**	$\frac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA}=\frac{\sum_i(p_i-\bar{p})(a_i-\bar{a})}{n-1}$, $S_P=\frac{\sum_i(p_i-\bar{p})^2}{n-1}$, $S_A=\frac{\sum_i(a_i-\bar{a})^2}{n-1}$

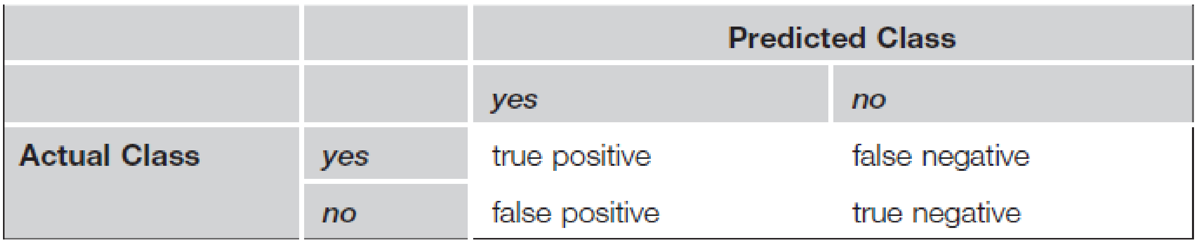
*Here, $\bar{a}$ is the mean value over the training data.
**Here, $\bar{a}$ is the mean value over the test data.

a) *Confusion Matrix*: In labor dataset example two classes are formed ie good & bad. The class good belong to the contracts which are accepted by the labor & management. And the contracts which are not accepted by the labor & management is under the bad class. In the two class case with classes bad and good ,it has four different outcomes. In confusion matrix two rows and two columns are made in which the entries are true positives (TP), true negatives (TN) , false positives (FP), false negatives (FN). The true positives (TP) are correctly classified ie it is actually positive values. A false positive (FP) is in reality negative value but by mistake it is represented in positive value. So false positive (FP) is wrongly estimated as positive values. A false negative (FN) is held when the outcomes in the table is in reality in positive but it is wrongly estimated as negative value. So the two class matrix is 2x2 form. The accuracy is the total number of correct classifications of true positives and true negatives divides by TP+TN+FP+FN.

Accuracy:

$$\frac{TP+TN}{TP+TN+FP+FN}$$

Table 4. Confusion matrix



		Predicted Class	
		yes	*no*
Actual Class	*yes*	true positive	false negative
	no	false positive	true negative

b) *Precision*: the true positive values are called precision. It is equal to the ratio of the true positive value to the total number of positive values in which the true positives and the false positive which are in reality is negative but wrongly estimated as positive values. It can be defined as:

$$Precision=\frac{TP}{TP+FP}$$

c) *False Positive Rate*: In false positive rate it is the division of false positive (FP) which is in reality is negative value but wrongly estimated as the positive values. It can be defined as:

$$\frac{FP}{FP+TN}$$

d) ***F-Measure:*** F-measure is the combination of the recall or true positive rate (TPR) and precision.

The formula for it is: $\frac{2*recall*\ precision}{Recall+\ precision}$

e) ROC (*Receiver operating characteristic*) curve: The ROC curve is represented in the graphs. The graph are represented in two dimensions in which horizontal axis is represented by false positives values. The FP values are in reality negative but wrongly estimated as positive value. And the vertical axis is represented with the value of true positives which are in reality is positive value. [6] They plot the true positive rate on the vertical axis against the false positive rate on the horizontal axis. The vertical axis ie true positives values in roc curves are expressed as a percentage.

f) ***Cross validation*****:** The cross validation is repeated in loops & by default the value of the repetitions is 10. In cross validation hold out method the training and testing data is taken. The testing data is used to correctly estimated the performance on the unknown data which are not given initially. In hold out technique the two third of the data is used for training in which the initial information is given already and the one third of the data is used for testing ie testing is used for correctly estimated the performance on the unseen data which is not given initially.

4. EXPERIMENT RESULTS AND PERFORMANCE EVALUATION

In our experiment we can done the categorization of the labor dataset ie classifies into the categories and then we can done the clustering of the dataset with the help of cobweb clustering algorithm. In this we can analyse the result of naïve bayes classifier and cobweb clustering algorithm and the results are accurate. The labor dataset are in arff format i.e attribute relation file format with all the attributes and it can contain class attribute with {bad, good} entries. So it can contain nominal attribute and missing values.

Table 5. Results of naïve bayes classifier

Tp rate	Fp rate	Precision	Recall	F-measures	MCC	ROC area	PRC area	Class
0.900	0.081	0.857	0.900	0.878	0.810	0.969	0.962	bad
0.919	0.100	0.944	0.919	0.932	0.810	0.969	0.980	good
Weighted average 0.912	0.093	0.914	0.912	0.913	0.810	0.969	0.974	

Table 6. Results of some error terms in naïve bayes classifier

Sr no	Terms	results
1.	Kappa statistic	0.8096
2.	Mean absolute error	0.1303
3.	Root mean squared error	0.2623
4.	Relative absolute error	28.4754%
5.	Root relative squared error	54.9395
6.	Coverage of cases(0.95 level)	98.2456%
7.	Mean rel. region size (0.95 level)	66.6667

Table 7. Results of correctly classified instances

Total no. of instances	57	
Correctly classified instances	52	91.2281%
Incorrectly classified instances	5	8.7719%

Table 8. Result of confusion matrix

	a	b
a=bad	18	2
B=good	3	34

Table 9. Result of cobweb cluster

No	Clustered instances	percentage
5	1	5%
9	2	10%
12	1	5%
13	2	10%
15	1	5%
17	1	5%
18	1	5%
19	2	10%
20	3	15%
23	2	10%
24	2	10%
25	2	10%

5. CONCLUSION

The text documents containing large amount of the data. On the web large or in high quantity of data is existed. So it is very difficult to classify & cluster the large amount of the data. The categorization of the data is to classify the text into predefined categories. The clustering is depend on initial factors for eg the k-means algorithm is used with the distance i.e the distance of the object is measured with the cluster centre. So in k means the distance is initial parameter. The clustering is depend on the text which is used for making clusters. Firstly select the clustering parameters very accurately for getting good results. More research is required for getting best results which are not complicated & attained quality. Some of the clustering algorithms can contain disadvantages for e.g when the clusters are formed that are not splited & it is unidirectional. So cobweb algorithm is come which is fast as compare to k- means algorithms. And cobweb is bidirectional algorithm. In clustering the one document can belong to more than one cluster so it is overlapped. So to deal with the problem considers some advanced algorithms which can allow overlapping of clusters. The research work can done on the areas of clustering, text classification, preprocessing of the data.

In the research work, taking the labor. arff (Attribute relation file format) dataset & firstly classified it with naïve bayes classifier & then clustered it with cobweb algorithm (Incremental hierarchical). Firstly preprocessing is done on the data i.e. cleaning of the data. A study of an integration of clustering and classification technique helps in identifying large data sets. The presented experiments show that integration of clustering and classification technique gives more accurate results. It can also be useful in developing rules when the data set is containing missing values. This integrated technique of clustering and classification gives a promising classification results with utmost accuracy rate and robustness.

REFERENCES

[1] George M. Marakas, Modern Data Warehousing, Mining, and Visualization, Pearson Education, New Delhi, 2005.

[2] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu , ―An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002.

[3] Sanjoy Dasgupta ―Performance guarantees for hierarchical clustering, Department of Computer Science and Engineering University of California, San Diego.

[4] Filkov and S. kiena. Integrating microarray data by consensus clustering. International Journal on Artificial Intelligence Tools, 13(4):863–880, 2004

[5] http://en.wikipedia.org/wiki/Naive_Bayes_classifier.

[6] Eibe frank & mark.a.hall "Data mining practical machine learning learning tool & techniques"