International Journal of Advanced Smart Convergence Vol.2 No.1 36-41 (2013)
DOI 10.7236/IJASC.2013.2.1.036

IJASC 13-2-10

Experimental Evaluation of Distance-based and Probability-based Clustering

Na Yeon Kwon[*], Jang Il Kim[*], Richard Dollein[**], Weon Joon Seo[***], Yong Gyu Jung[*†]

[*]Dept. of Medical IT Marketing, Eulji University, 461-713 Korea
[**]R&D Center, Softforum Co., LTD, 9th FL., Hancom Tower, 463-400 Korea
[***] R&D Center, Softforum Co., LTD, 9th FL., Hancom Tower, 463-400 Korea

Abstract

Decision-making is to extract information that can be executed in the future, it refers to the process of discovering a new data model that is induced in the data. In other words, it is to find out the information to peel off to find the vein to catch the relationship between the hidden patterns in data. The information found here, is a process of finding the relationship between the useful patterns by applying modeling techniques and sophisticated statistical analysis of the data. It is called data mining which is a key technology for marketing database. Therefore, research for cluster analysis of the current is performed actively, which is capable of extracting information on the basis of the large data set without a clear criterion. The EM and K-means methods are used a lot in particular, how the result values of evaluating are come out in experiments, which are depending on the size of the data by the type of distance-based and probability-based data analysis.

Key words: EM, Decision-making, K-Means, Maximization Step, BI-RADS assessment

1. INTRODUCTION

Computers and Internet evolves rapidly, all organizations and companies have decided to build a database to recognize as information infrastructure data. By becomes enormous size as finding a simple query tool is difficult to grasp the characteristics, the database was able to obtain a new knowledge could not know until now. The patterns and relationships are inherent in the database, and to provide the necessary information to decision-making, but the cluster analysis and to find a pattern in small groups with similar characteristics a variable of this data is called. And cluster analysis in the characteristics of the data of is often used to find a group of similar items. Therefore, in this paper, we examine what the EM technology commonly used in the cluster analysis, let's analyze how the value of the result of the EM method or out depending on the size of the data.

Manuscript received: Mar. 26, 2013 / revised : Apr. 13, 2013
Corresponding Author: ygjung@eulji.ac.kr
Tel: +82-31-740-7190, Fax: +82-31-740-7190
Dept. of Medical IT Marketing, Eulji University, Korea

2. RELATED RESEARCH

2.1 Counting the cost

Two confusion matrix for three classes ABC in Fig. 1.The table on the left, is the case of the actual forecast, table on the right, to the prediction with random values.In the test set a total of 200 pieces, the success rate will be 70% Become a 140 with a 88 +40 +12 cost of diagonal if the results match predictions and the actual success rate of the actual forecast, but right If I show a success rate lower than 82 units when the prediction of success. To extract information without separate class when extracting any information from the database and finally also shows that likely will be extracted after the prediction.

		Predicted class			
		a	*b*	*c*	*total*
Actual class	*a*	88	10	2	*100*
	b	14	40	6	*60*
	c	18	10	12	*40*
	total	*120*	*60*	*20*	

		Predicted class			
		a	*b*	*c*	*total*
Actual class	*a*	60	30	10	*100*
	b	36	18	6	*60*
	c	24	12	4	*40*
	total	*120*	*60*	*20*	

Fig. 1. Different outcomes of a three-class prediction

2.2 Concept of the EM

EM algorithm is a congestion algorithm first proposed by Hartley in 1958, was organized by Dempster in 1977. Similar to the K-Means algorithm, after you create the initial model, EM algorithm will continue to create a model with optimized model through the purification process repeated. EM algorithm will continue to create a model of optimum by adjusting possibility via the repeated purification process; each object belongs to a mixture model to probability. The K-Means algorithm is using Euclidean distance function by the log-likelihood function, while EM algorithm to evaluate the suitability of the model. In other words, K-Means congestion is a method of distance-based, while EM Probability-based clustering.

K-Means algorithm used a similar approach and virtually in the congestion algorithm belonging to the method of division. However, unlike to evaluate goodness of fit with the distance K-Means, EM is the major difference is that it evaluates the goodness of fit with that probability. Congestion probability-based, using the mixture model for the distribution of data. Congestion means the data distribution of one here. The congestion probability-based, it can record one belongs to more than one model. Extent to which belong at this time is to be given as probability weight.

3. EXPERIMENTS

3.1 EM Algorithm

In Fig. 2, the EM algorithm is the number of clusters, k and termination conditions before performing. The algorithm is divided into M-Step and E-Step large. Calculate the probability that the probability distribution for the k, each record belongs, and assign it to convert the weight E-Step is Expectation Step. M-Step next Maximization Step updates the model. In order to assume a normal distribution, this process changes the mean and standard deviation of records assigned new models from EM. At this time, because it has a weight element about each record, it will be calculated taking into account the weights.

Input : Cluster number k, a database, Stopping tolerance ε (> 0)
Output : A set of k clusters with weight that maximize Log-likelihood function.
(1) Expectation Step
For each database record x,
Compute the membership probability of x in each cluster h = 1,...,k.
(2) Maximization Step
Update mixture model parameter (probability weight)
(3) Stopping criteria
If stop criteria is satisfied stop
Else set j = j+1 and goto (1)

Fig. 2. EM Algorithm

3.2 Experiment

The data in Fig. 3 is a data consisting of results of four subjects mathematics of 10 students, physics, English, and Chinese. [Mathematics, physics] is a good student, three students of the first, is a good student [English, Chinese] the three following. 2 people are an excellent student all four subjects, a student not all subjects of all, that is, and the last two, is a simple data that is configured on a herd of four. Therefore, it is cast WEKA program data.

	A	B	C	D
1	math	physics	english	chinese
2	90	100	40	50
3	80	90	30	45
4	85	95	50	40
5	30	50	100	90
6	50	40	90	95
7	35	45	95	100
8	100	90	95	100
9	87	95	95	90
10	10	40	30	50
11	50	30	20	60

Fig. 3. Student performance data

```
@relation student_score
@attribute math numeric
@attribute physics numeric
@attribute english numeric
@attribute chinese numeric
@data
90,100,40,50
80,90,30,45
85,95,50,40
30,50,100,90
50,40,90,95
35,45,95,100
100,90,95,100
87,95,95,90
10,40,30,50
50,30,20,60
```

Fig. 4. Format Arff data

WEKA is a data mining program that was developed in the Java language. Unlike what not to contact data mining programs plurality expensive, it can be obtained easily by program provided free of charge. The system provides source code for the Java language of the entire program as open source other than anything; it is a very useful program developer to develop data mining application to reference. Because it is a program that is not a commercial program was produced for research purposes, it is possible to feel some difficulty when using. It is free, offers analysis algorithms many different rather than pay some programs, analyzing visual features is also excellent and it is a program that can be effectively used to analyze the data. For taking the WEKA program, the data type of the above-mentioned. Use the Arff format as shown in Fig. 4

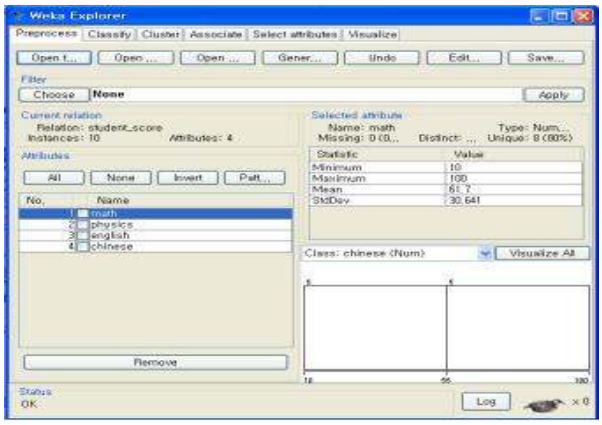

Fig. 5. WEKA program

Turning to Fig. 5, the variable name display four are middle left, the screen now talking selected variables, Selected attribute upper right is selected mathematically. In the table below, statistics of the selected variables (minimum value, maximum value, average value, deviation) are out. So it will be moved to cluster tab is a third tab to the cluster analysis on the basis of this.

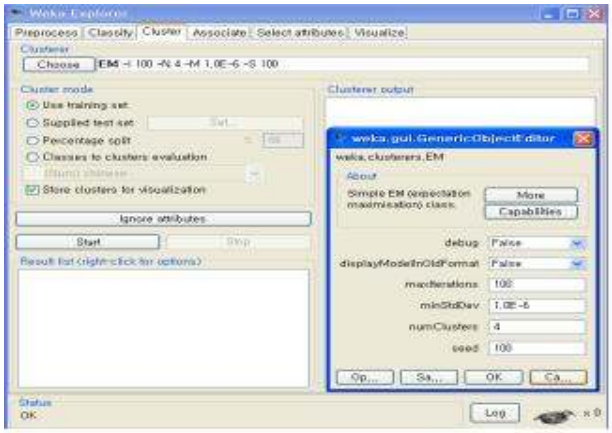

Fig. 6. EM algorithm selects Options

Though it is a screen that comes out when you click the Cluster, EM algorithm is selected basically already, Fig. 6, can also be accomplished by selecting the other algorithms, to change the option. When you click the Choose button, it will be displayed for you to select the algorithm. After you click the text EM, if you click on the empty space of the screen, a window where you can change the options will be displayed. The called numClusters in the Options window is that it refers to the number of crowded, fill in the four the number of crowded we can crowded, and assumed to be congestion of four, and we have created a data I experiment

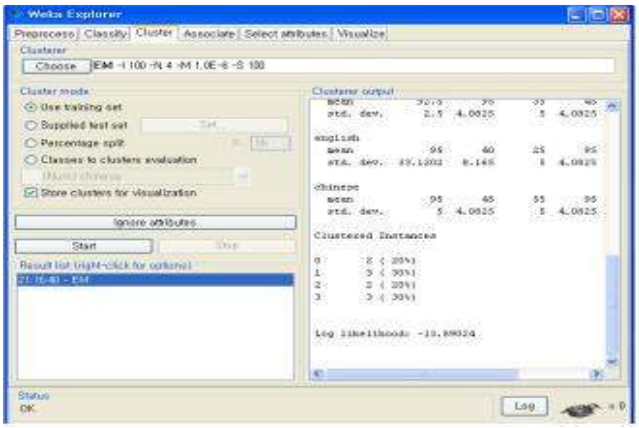

Fig. 7. EM algorithm

Looking at Fig. 7, looking at the bottom of the right pane, it can be seen that it has been created congested four. So name congestion 0, 1 congestion, congestion 2, 3 of congestion has been granted. First, it is possible to grasp the number of records contained per congestion. Three students belong to the congestion 0.The congestion 2, 3 to congestion, the congestion 4 of 2 people, it is out on the results of three people and belongs to two people is not belong in this way.

Attribute	Cluster 0 (0.2)	1 (0.3)	2 (0.2)	3 (0.3)
math				
mean	93.5	85	30	38.3333
std. dev.	6.5	4.0825	20	8.4984
physics				
mean	92.5	95	35	45
std. dev.	2.5	4.0825	5	4.0825
english				
mean	95	40	25	95
std. dev.	33.1202	8.165	5	4.0825
chinese				
mean	95	45	55	95
std. dev.	5	4.0825	5	4.0825

Fig. 8. EM algorithm data

Fig. 8 is a diagram startup screen of Fig. 7, an enlarged clearance of congestion. In the Cluster part, it is a summary of the first cluster 0 congestion. 20% of the total, i.e., 0.2, is a sense that contains the data of one 2 probability though to talk of numbers that are on the bottom of the 0. Information about the variable (average, deviation) is displayed directly below it. 95 average, deviation 33, 95 average, China is a deviation 5.0 92.5 average, deviation 2.5, English 93.5 average, deviation 6.5, science and math. In other words, congestion 0 is possible to understand that math, science, English, Chinese, which is an excellent group performance of all. Mathematics, science is superior run Chinese English is a group

performance is low, the Cluster2, Cluster1 is a group performance of all subjects is low, The Cluster3, grades mathematics, science is low, indicating that it is a group results in English and Chinese high.

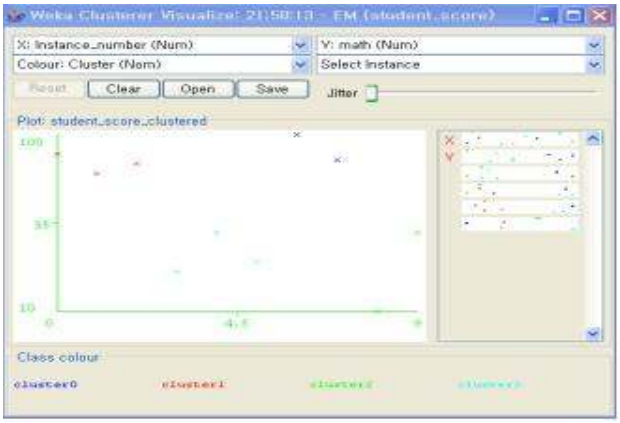

Fig. 9. Chart for EM algorithm

If you look at the (XY graph) white part of the center, asterisk (Show 10 students) 10 records (*) display has been. X-axis current is a classification by number of congestion, Y axis is what you view the score of mathematics. Each asterisk (*), can be viewed to determine the records contained in the congestion of each other that they are displayed in different colors depending on the congestion.

This will analyze large amounts of data via the EM algorithm. Mammographic Mass Data BI-Severity RADS assessment, Age, Shape, Margin, Density, consists of a total of 6, and has a total of 961 instances. Therefore, because it is data that will determine the Severity see BI-RADS assessment, Age, Shape, Margin, and Density, looking at the attribute, the BI-RADS assessment, is represented by a number between 0-6, this data, Age is, the Shape, the 1 ~ 4, Margin, 1 ~ 5, Density is represented by 1 ~ 4, Severity will us decide whether 0 or 1 through them by the exact age respectively.

	A	B	C	D	E	F
1	BI-RADS assessment	Age	Shape	Margin	Density	Severity
2	5	67	3	5	3	1
3	4	43	1	1	?	1
4	5	58	4	5	3	1
5	4	28	1	1	3	0
6	5	74	1	5	?	1
7	4	65	1	?	3	0
8	4	70	?	?	3	0
9	5	42	1	?	3	0
10	5	57	1	5	3	1
11	5	60	?	5	1	1
12	5	76	1	4	3	1
13	3	42	2	1	3	1
14	4	64	1	?	3	0
15	4	36	3	1	2	0
16	4	60	2	1	2	0
17	4	54	1	1	3	0
18	3	52	3	4	3	0
19	4	59	2	1	3	1
20	4	54	1	1	3	1
21	4	40	1	?	?	0
22	?	66	?	?	1	1

Fig. 10. Mammographic Mass data

4. EXPERIMENTAL RESULTS

Fig.11, 12, and 13 through the WEKA program is the result of analysis of the EM algorithm, the data on the analysis of the two communities. Looking at them, the BI-RADS assessment, 50 congestion 1 in 0 crowded out 60 about Age it is possible to see congestion 0 that there is no significant impact in the cluster analysis of four crowded one came out in 4,7 crowded one will be able to see that age is good to have. But much higher compared to cluster 1 and cluster levels 0, 1,2 In Shape or similar in the three figures, has much higher contrast, the 4 lower in cluster 0 and cluster 1 that can be seen. Once you have the numbers 1, 2, and go to 0 communities only when you Shape is equal to the numerical value of 4, go to the communities. In Fig. 4, 5 Margin Looking at figs. 1 0 communities is much higher, the when similar Fig. 2,3.

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100
Relation:     mammographic_masses
Instances:    961
Attributes:   6
              BI-RADS_assessment
              Age
              Shape
              Margin
              Density
              Severity
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


EM
==

Number of clusters selected by cross validation: 2
```

Fig. 11. EM algorithm analysis

```
Attribute                       0          1
                             (0.52)     (0.48)
=============================================
BI-RADS_assessment
  mean                          4     4.7321
  std. dev.                1.7812      2.526

Age
  mean                    50.5576    60.9197
  std. dev.               13.5709    13.3692

Shape
  1                      184.7227    41.2773
  2                      173.7422    39.2578
  3                       54.8437    42.1563
  4                       94.4894  338.5106
  [total]                 507.798    461.202
Margin
  1                      343.7315    63.2685
  2                       12.2597    13.7403
  3                       45.5092    72.4908
  4                       92.2727  189.7273
  5                       15.0248  122.9752
  [total]                 508.798    462.202
```

Fig. 12. EM algorithm Attribute

Density, all numerical values are, on average, when there is no error. So when I saw this, Density means that there is a big impact in the cluster analysis. Finally, looking at the Severity,

value 0, 0 congestion is much higher, the number 1, the result is high congestion 1 is out. Congestion and congestion 1 0 approximate data has not been changed completely clustering, is shown clustering has become well.

```
Density
  1                    8.9763   9.0237
  2                   37.2758  23.7242
  3                  458.2555 417.7445
  4                    3.2904  10.7096
  [total]            507.798  461.202
Severity
  0                  425.4066  92.5934
  1                   80.3914 366.6086
  [total]            505.798  459.202
Clustered Instances

0      493 ( 51%)
1      468 ( 49%)


Log likelihood: -9.18885
```

Fig. 13. EM algorithm Density

5. CONCLUSIONS

In recent years, efforts for using data mining methods in each field to obtain the information is underway, no standards specifically mentioned, algorithms cluster analyzing classification is able to apply the EM algorithm tried to. The analyzed data of different sizes by applying the EM algorithm of WEKA program, but it was crowded full as predicted based on the performance data of students who experiment first. Result of the experiment, there is some errors but Mammographic Mass data was seen that it has been group of Jip through some numbers low high. That is, when the cluster analysis data using the EM algorithm blowing also be a difference in size of the data, clustering, the data indicate that the No significant errors. However, it was found in this experiment that the speed is slow as the size increases, in this paper, we made a comparative analysis in accordance with the size of the data passed through only the EM algorithm in the method of congestion, but the future consider the scheme that can be to try to comparative analysis of congestion various methods also analyze the EM algorithm, and overcrowd the large amount of data in a short time.

REFERENCES

[1] Stuart Moran, Yulan Hey, Kecheng Liu, “An Empirical Framework for Automatically Selecting the Best Bayesian Classifier”, Proceedings of the World Congress on Engineering 2009 Vol I, WCE 2009, July 1 - 3, 2009.

[2] Carolina Ruiz, “Illustration of the K2 Algorithm for Learning Bayes Net Structures”, Department of Computer Science, WPI, 2005.

[3] EvelinaLamma, FabrizioRiguzzi, Sergio Storari, “Improving the K2 Algorithm Using Association Rule Parameters”, Modern Information Processing: From Theory to Applications B, 2006.

[4] Jesse Davis and Pedro Domingos (2010). Bottom-Up Learning of Markov Network Structure. In the *Proceedings of the 27th International Conference on Machine Learning (ICML)*.

[5] JesúsCerquides, “Tractable Bayesian Learning of Tree Augmented Naive Bayes Classifiers”, Ramon López de Màntaras, 2003.

[6] Yong Gyu Jung, Jong Han Lim, Automobile Traffic Accidents Prediction Model using by Artificial Neural Networks, ICHIT2012, Communications in Computer and Information Science, Vol.310,p713-p719

[7] Yong Gyu Jung, Song Ei Han Ranking Methods of Web Search using Genetic Algorithm, Journal of the Institute of Webcasting, Internet Television and Telecommunication (IWIT), Vol.10 No.3 p91-p96, June 2010

[8] Yong Gyu Jung, Go Eun Heo, Ensemble Classification Method for Efficient Medical Diagnostic, Journal of the Institute of Webcasting, Internet Television and Telecommunication (IWIT), Vol.10. No.3 p97-p102, June 2010

Yong Gyu Jung
received the B.S. in physics Education from Seoul National University in 1981. And then he got the M.S. and ph.D.degreeofComputer Science from Yonsei and KyonggiUniversityin 1994 and 2003 respectively. Since 1999, he has been a Faculty of Medical IT marketing Dept. in Eulji University. His research interests are in the areas of medical information analysis and international standards including e-Business. He is a Member of the UN/ECE/CEFACT and ISO/TC154 standard organization.

Na Yeon Kwon
performanced comparison of ID3 and C4.5 decision tree algorithms, such as research and enrolled in the Department of Medical ITMarketing, Eulji University. OCS, EMR in hospitals and medical information systems implementation technology and are interested in data mining technique for the analysis of clinical data.

Jang Il Kim
obtained a B.S. in physics from the Suncheon National University and now a graduate student at the Department of IT Marketing in Eulji University.He is a member of KISA phishing Center Advisory Consultant and interested in the field of security and medical information

Weon Joon Seo
received the B.S. in Information Engineering in 1996 and the M.S. degree of Information Coummunication Engineering from Sungkyunkwan University(SKKU) in 2001respectively. He has worked on security companyof Softforum Co., LTD. He is interested in the areas of security and analysis of big data.

Richard Dollein
has been a CEO of DÖLLOMATIK at Burgebrach, which is currently the only specialized craft operating in Bavaria, in the field of manufacturing and marketing automatic sliding door systems with their own development. He is interested in the areas of automatic control and security in the bank and airport fields