Volume : 3 | Issue : 6 | June 2014 • ISSN No 2277 - 8160



Research Paper | Computer Science

Customizing Clustering Algorithm for Data Mining for Lead Generation

E. Manigandan Research Scholar, SCSVMV University, Enathur, Kanchipuram-631561

Dr.V.Shanthi Professor, Dept. of MCA., St. Joseph College of Engineering, Chennai-119.

Magesh Kasthuri Research Scholar, SCSVMV University, Enathur, Kanchipuram-631561

ABSTRACT

Lead generation is a technique used to attract or generate customer leads using lead generation techniques like contextual advertising (to be discussed below). It is an easy and painless way of attracting people/users and cultivating prospective customers out of them.

This study is taken to customize Clustering Algorithm for Data mining process in a Customer Relationship Management field as the space of data is more and variant [17]. Also proving the usability of customized Clustering Algorithm in Data Mining and the efficiency in processing mechanism compared to other costlier methods used in current situation of Data mining in Customer Relationship Management is part of this study.

Clustering Algorithm is broadly used as Spectral algorithm in Social media [2], where a reference of a contact is used and mined further for various combinations of suggested friends and lookups.

KEYWORDS : Data Mining, Clustering Algorithm, Spectral, Spacial algorithm, Classification, KMeans.

I. Introduction

While various studies have recorded the favorable circumstances and depicted the information mining process, there is research on the impact of Lead era regarding Data mining Process. Presently with the appearance of Bigdata dissection and information purging for Customer Relationship Management, Lead era calculations are getting more centered and refined as far as transforming quality comes about.

Lead sustaining and lead era procedures are next center regions as far as arranging information for digging the effects needed for Customer Relationship Management reports. The center of Spatial and ghastly calculations might assume an indispensable part in lead era for current drift in engineering viz social figuring and clients utilizing interpersonal interaction.

The customized algorithm developed as part of this study considers two major areas of data mining using clusters. This algorithm is named as M-Cluster Algorithm.

Marketing a product is an important area to introduce, popularize, improve and sustain business. The marketing technology grows with the growth of technology and industrial changes. Earlier Marketing was done with Banners and Advertisement hoardings which are then grown to TV broadcasts, online advertising and now with the growth and level of usage of Internet; Marketing is in a peak stage of booming ideas like Lead-gen and contextual advertising.

Earlier, marketing people will take survey in roads, buses and then in road shows which is then matured to stalls in symposium, shows and tele-marketing and chain marketing. The advantages of these techniques are numerous but with the advent of online shopping and internet usage, this technique has gone to its next dimension called Lead generation.

By concept, Lead generation is simple to understand but difficult and swirls when tries to implement it practically. When we understand the types of lead generation and the techniques that are popularly used for lead generation, we will get a fair idea of Lead generation concepts and methodologies.

- To Identify key Lead generation techniques to be used in Customer Relationship Management for sales marketing to decide data gathering.
- To define key merits and demerits of these techniques and prepare a Matrix of comparison of these techniques to justify the data source and data set.
- To Use M-cluster Algorithm for Lead qualification (training set preparation and data evaluation)
- To Prepare Case studies using this Algorithm to qualify Leads for Search engine based Customer Relationship Management
- To Define steps in Cost effectiveness in Lead qualification (cost per lead) in search engine
- To Define an Analytics solution using Drupal based Customer Relationship Management for Lead qualification output from M-Cluster
- To prepare an algorithm integrated in WEKA to evaluate experimental result.

II. Experimental setup

Contextual advertising is a form of targeted advertising (advertisements targeted only to intended users) for advertisements appearing on websites like advertisements displayed on top and right of yahoo mail service. The advertisements themselves are selected and served by automated systems based on the content displayed to the user and thus called contextual.

Now we come to the basic and core need of the topic in discussion-business. As we come through different areas of lead-gen, we will be eager to know how business is generated with lead-gen. lead nurturing is the vital part of lead-gen which concentrates required information to the advertisers from the leads generated.

Lead-gen is mainly conceptualized for better Return Of Investment (ROI) through easy and meaningful way of marketing. Revenue can be generated from lead-gen in four broad categories as follows:

- Cost per Advertisement
- Cost per Impression
- Cost per click (online)
- Cost per lead

Since MCluster algorithm is using Ranking based samples as compared to Best first based search in KMeans or other most commonly used data mining algorithms, there is a definite benefit in MCluster when used with continuous growing set of data and also in improvised time taken for processing the records.

Techtarget (www.techtarget.com) is one of the largest lead-gen providers in IT industry. It has as many as 41 inter websites (theserverside, searchSOA etc.,) which publishes contents freely to the users. Registered users can freely browse and download interested articles from this site. The word 'Registered' is highlighted in previous sentence as lead-gen starts from there in Techtarget where certain data

from user is taken and sold to advertisers based on the interest (lead nurturing).

Once the leads are gathered from one or some of the above techniques, we have the raw leads ready to be processed and distributed to advertisers. They can be processed manually or using Customer Relationship Management tools. When we process the raw leads manually, it involves lot of experience to the processing team to carefully analyze and generate leads out of it. When the volume of leads are more or when the information to be processed in each lead is more and complex (meshed information between different data), then manual processing will be a cumbersome job and can result into failure results of preparing wrong leads or missed leads.

The case study taken for evaluation is on improvising sales leads in computer world site (a sample website for IT communication and trends). Example: techtarget.com

The data sources are collected from various sources of computer world using various techniques including questionnaire and data extraction techniques (using drupal).

MCluster is implemented in an experimental evaluation setup as a WEKA attributeSelection class and Cluster classifier class to validate the processing time/results from MCluster evaluated data.

For example, when a growing set of records (in both dimensions of fields and records) is fed to MCluster algorithm, we found the consistency in processing rate and time execution as shown below:

Fields	Records	Total Data (Fields X Records)	Time taken with MCluster algorithm	
			To build Clusters (in ms)	Attribute selection (in ms)
5	14	70	62	78
5	24	120	39	15
5	150	750	16	234
17	57	969	23	110
7	209	1463	47	131
8	209	1672	63	167
36	683	24588	453	406
20	810	16200	437	188
9	768	6912	78	125
21	1000	21000	250	25
20	1500	30000	1093	328
9	12960	116640	1094	151
15	32561	488415	1753	6188

As a sample, when a dataset like below is taken and processed for cluster generation in MCluster

III. Comparision with kmeans

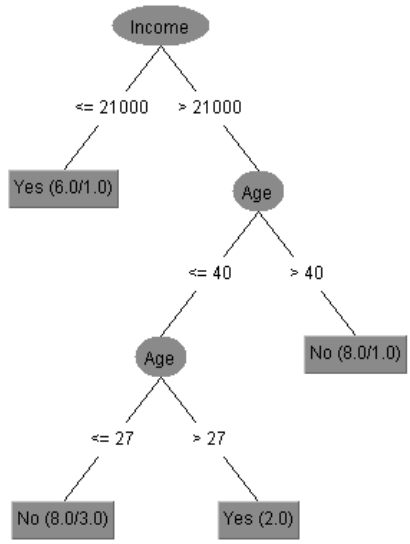
Data set selection should be either random or based on historical data processing results. Hence, this works in a supervised methodology.

The data storage (data model used to store the data set for filtering and processing) is not as efficient as they are not indexed on a matrix representation of multi-folded field representation.

Algorithm customization is not so generic or flexible solution as they re-create the processing cycle from the beginning. Hence it is a costlier process to run various iteration/cycles of the processing to refine the result ratio or reduce error rate.

To overcome all these, M-Cluster is customized in such a way that:

- Store the data model in our own clustering representation in a matrix format associating each unique field in an indexed searchable fashion.
- Runs in unsupervised methodology by storing various fields in indexed model and based on the result required, the field selection and its corresponding test set selection would happen.



Sample Data showing clustered result

M-cluster algorithm can be customized in such a way that it can run the training set cycle multiple times (and hence cost efficient as the cycle runs in a random selected smaller data units only) to refine the result ratio and error rate

KMeans is a classification algorithm that require the data to be normalized (k) so that no particular variable or subset of variables dominates the analysis. KMeans doesn't use training set for data model preparation and K-means Clusters should have small within-cluster variation compared to the between cluster Variation. The proposed algorithm is a historical research methodology whereas M-cluster algorithm follows field research methodology.

Fields	Records	Total Data (Fields X Records)	Time taken with K-Means (in ms)	Time taken with MCluster (in ms)
5	14	70	100	62
5	24	120	32	39
5	150	750	16	16
17	57	969	31	23
7	209	1463	124	47
8	209	1672	47	63
36	683	24588	453	453
20	810	16200	406	437
9	768	6912	94	78
21	1000	21000	266	250
20	1500	30000	1110	1093
9	12960	116640	1141	1094
15	32561	488415	1787	1753

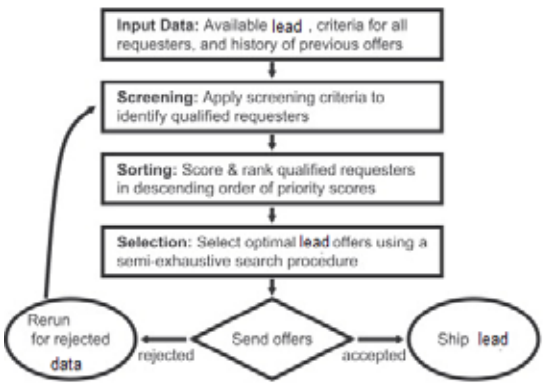
There is a good improvement in MCluster compared to KMeans in terms of processing time when the data is scaling up to big data as shown below:



Record process vs Time taken

IV. Scope of Evaluation

- Evaluation of the M-Cluster algorithm and comparison with other standard Clustering algorithm with respect to cost effectiveness, efficiency, improvisation, usability and adoptability.
- Sample case study evaluation using input source collected from techtarget.com (as a sample) and evaluation of M-cluster including training set identification.

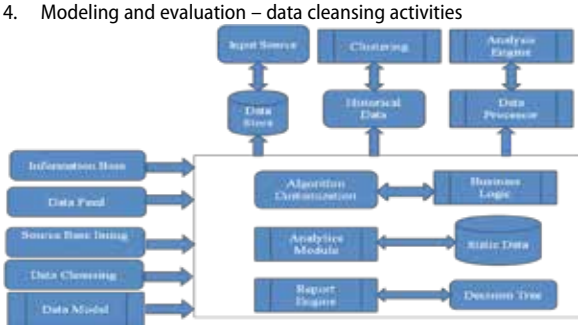


Lead generation and Data mining process

- Benchmark data evaluation and statistics preparation on performance speed, data handling efficiency to name a few.
- To study the factors improvised by M-cluster algorithm in CRM data mining.

High level Steps involved in the system are:

1. Requirement Understanding and base lining
2. Data Set definition and boundary condition defining
3. Data preparation and collection from various sources
4. Modeling and evaluation – data cleansing activities



5. Analysis and Refining the data – usage of clustering algorithm to refine the data into groups and hierarchy of clusters
6. Inference and report generation

The algorithm steps is explained as below:

Step-1: From a given Input set (s), prepare group each set of elements (of likely group from predictive column) into a cluster of elements

Step-1.1: Get user input on error rate expected € [tolerance level] and input base lining T (threshold)

Step-2: Identify the size of training set based in T and € [TS = I * € / T] where I is input volume (size).

Step-2.1: From each cluster, pick a set of elements (having relation) and examine if it can be used for predictive results (identification of training set) based on €

Step-3: If the training set is not effective, repeat step-2.1 until a predictive training set is identified

Step-4: For each cluster (Q)

Step-4.1: Remove the top element of Q (say u) and merge it with its closest cluster U.closest (say v) and compute the new representative points for the merged cluster w.

Step-4.2: Also remove U and v from T and Q.

Step-4.3: For all the clusters x in Q, update x.closest and relocate x (this is to update the address reference for relative child/node element in the cluster)

Step-4.4: Insert w into Q

Step-4.5: Repeat Step 4

Step-5: Traverse each cluster for the mean elements and prepare results based on the means of evaluation (M-cluster evaluation).

V. Summary

This study helps in proving the stability of the processing algorithm and also the efficiency in processing mechanism compared to other costlier methods used in current situation of Data mining in CRM.

A sample output showing statistics of evaluation (when run in WEKA)

```
Attribute Evaluator (supervised, Class (nominal): 5 Health Insured):
        Fano Ranking Filter
5 Columns X 24 Nodes. Time to Evaluate Cluster data based on trained set: 453 ms

Ranked attributes:
 2.178  4 Education
 1.836  2 Gender
 0      1 Age
 0      3 Income

Selected attributes: 4,2,1,3 : 4
```

We took data analysis of technical users visiting various forums like technical, musical, social, cultural, historical interests and did a Data mining and analytics processing to gather customer information based on regional and cultural backgrounds. The results produced tremendous trends that the clustered algorithm are suitable for any data mining process when scaling and data classification are diversified and less in control.

REFERENCES

Example sources for dataset for evaluation: | http://repository.seasr.org/Datasets/UCI/arff/ | https://dreamtolearn.com/doc/2HDNJH3X-JU6CVGKZ7SDM4MCSW | http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html | http://kdd.ics.uci.edu/summary.data.application.html | http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets | [1] Yuqiang Fang, Ruili Wang and Bin Dai, "Graph-oriented Learning via Automatic Group Sparsity for Data Analysis", Data Mining (ICDM), 2012 IEEE 12th International Conference. pp. 251 – 259 | [2] Robert Robert Jenssen, Torbjørn Eltoft and Jose C. Principe, "Information Theoretic Spectral Clustering", 2004 IEEE,Vol.1. 0-7803-8359-1. pp.111-116 | [3] osama Abu Abbas, "Comparisons Between Data Clustering Algorithms", the international arab journal of Information Technology, vol.5, No.3, july 2008. | [4] Hsin-Chien Huang, Yung-Yu Chuang and Chu-Song Chen, "MULTI-AFFINITY SPECTRAL CLUSTERING" – International Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference, pp. 2089 – 2092 | [5] Guo Wensheng and Li Guohe, "Text Clustering Algorithm Based on Spectral Graph Seriation" - Control and Decision Conference, 2009. CCDC '09. Chinese, pp. 4255 – 4259 | [6] Robles-Kelly, A., Hancock, E.R. "Graph edit distance from spectral seriation" - Pattern Analysis and Machine Intelligence, IEEE Transactions on (Volume:27 , Issue: 3) pp. 365 - 378 | | [7] Antonio Robles - kelly, Edwin R. Hancock, "Graph edit distance from spectral seriation" - IEEE International Conference on Computer Vision(ICCV 2003) 2-Volume Set, 0-7695-1950-4/03 pp.65-78 | [8] Antonio Robles-Kelly, Edwin R. Hancock , "Graph Matching Using Spectral Seriation and String Edit Distance", - Lecture Notes in Computer Science Volume 2726, 2003, pp. 154-165 | [9] Hang Yu,Edwin R. Hancock, "Graph Seriation Using Semi-definite Programming" - Graph-Based Representations in Pattern Recognition Lecture Notes in Computer Science Volume 3434, 2005, pp. 63-71 | [10] Ye Xiao-rong, "Analysis on Network Clustering Algorithm of Data Mining Methods Based on Rough Set Theory" - Knowledge Acquisition and Modeling (KAM), 2011 Fourth International Symposium, pp. 296 – 298 | [11] Peng Jin ,Yun-Long Zhu and Kun-Yuan Hu, "A Clustering Algorithm for Data Mining Based on Swarm Intelligence" - Machine Learning and Cybernetics, 2007 International Conference on (Volume:2) pp. 803-807 | [12] Zhijie Xu, Laisheng Wang, Jiancheng Luo and Jianqin Zhang, "A modified clustering algorithm for data mining" - Geoscience and Remote Sensing Symposium, 2005. IGARSS '05. Proceedings. 2005 IEEE International (Volume:2) pp. 155-158 | [13] Tianzhen Wang, Tianhao Tang, "A New Data Mining Method based on Fusion Clustering Algorithm", 2005. ICNN&B '05. International Conference on (Volume:2) pp. 706 – 711 | [14] Chapman and Hall , "Data Clustering: Algorithms and Applications",CRC. 1 edition (August 21, 2013) ISBN-10: 1466558210 pp. 13-19 | [15] Dehzangi. O, Zolghadri, M.J., Taheri, S. and Fakhrahmad, S.M, "Efficient Fuzzy Rule Generation: A New Approach Using Data Mining Principles and Rule Weighting Fuzzy Systems and Knowledge Discovery", 2007. FSKD 2007. Fourth International Conference on (Volume:2) pp.134-139 | [16] Ramakrishnan, G., Joshi, S., Negi, S., Krishnapuram, R "Automatic Sales Lead Generation from Web Data - Data Engineering", 2006. ICDE '06. Proceedings of the 22nd International Conference pp.101-103 |