Available online at www.elixirpublishers.com (Elixir International Journal)

Information Technology

Elixir Inform. Tech. 91 (2016) 38047-38050

ISSN: 2229-712X

Spectral Clustering in Data mining with the Case Study of Customer Relationship Management

E.Manigandan[1], V.Shanthi[2] and Magesh Kasthuri[1]

[1]Research Scholar, SCSVMV University, Enathur, Kanchipuram-631561.

[2] Professor, Department of MCA, St. Joseph's College of Engineering, Chennai-119.

ARTICLE INFO

Article history:

Received: 23 December 2015;

Received in revised form:

28 January 2016;

Accepted: 2 February 2016;

Keywords

Data Mining,
Clustering Algorithm,
Spectral,
Spacial Algorithm,
Classification,
KMeans.

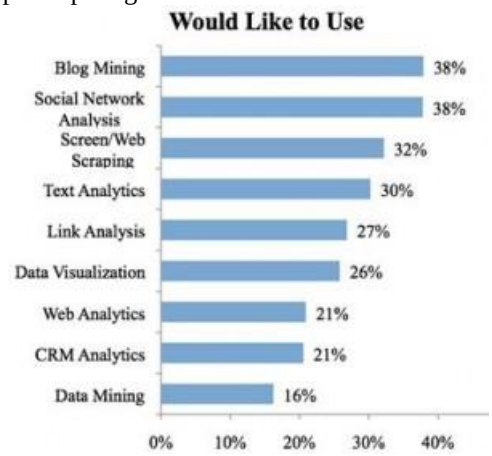
ABSTRACT

In Data mining world, Lead generation is a data searching technique which is used to collect relevant customer information (leads), one of the examples for this techniques is contextual advertising. You might have noticed as soon as you open google site to search something, it displays unique advertisement or sponsored link along with search results. This sponsored link is typically based on search text, user logged in (ex: google user), location, browser to name a few. This type of preparing customized advertisement and sponsored links is called as Contextual advertisement and this technique is an example for Lead generation. It is an easy and painless way of attracting people/users and cultivating prospective customers out of them. The key idea of this paper is to bring out the importance of data mining in the field of CRM and also to explain the benefits of M-Clustering algorithm which we propose for data mining which proves to be efficient as it uses clustering approach compared to k-means algorithm. Also, there is a comparison with Newman's algorithm where the significance is highlighted in terms of training set and historical data handling in M-Clustering.

Introduction

Large complex graphs representing relationships among sets of entities are an increasingly common focus of scientific inquiry. Examples include social networks, Web graphs, telecommunication networks, semantic networks, and biological networks. One of the key questions in understanding such data is "How many communities are there and what are the community memberships"?

There was an Integrated marketing survey done in 2013 on analysing marketing solutions in Customer Relations and the survey outcomes proves that there are a large volume of unwanted or irrelevant data passing to the customers where either they are not interested or not showing interested in actively participating.



Integrated Marketing survey (Source: http://www.web-cubed.com)

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data is called Knowledge discovery.

Problem description

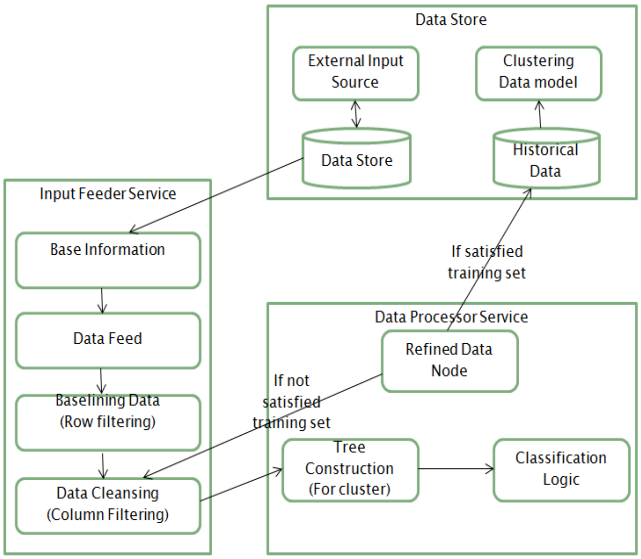
Algorithms for finding such communities, or automatically grouping nodes in a graph into clusters, have been developed in a variety of different areas, including VLSI design, parallel computing, computer vision, social networks, and more recently in machine learning. Good algorithms for graph clustering hinge on the quality of the objective function being used. A variety of different objective functions and clustering algorithms have been proposed for this problem, ranging from hierarchical clustering to max-row/min-cut methods to methods based on truncating the Eigen space of a suitably defined matrix. In recent years, much attention has been paid to spectral clustering algorithms.

Study of Literature

Newman [10] and Newman and Girvan [9] showed across a wide variety of simulated and real-world graphs that larger Q values are correlated with better graph clustering. In addition, they found that real-world un-weighted networks with high community structure generally have Q values within a range from 0.3 to 0.7. By comparing the new algorithm (M-Cluster) with Newman's algorithm on different graph data sets and empirically illustrate that the spectral approach to maximizing Q produces results that, in terms of cluster quality, are comparable or better than results from Newman's hierarchical algorithm, and the proposed algorithms are linear

Tele:
E-mail addresses: ibmmani78@yahoo.com

per iteration in the number of nodes and edges in the graph, compared to quadratic complexity in the number of nodes for the original algorithm proposed by Newman.



Tasks involved in Clustered Analysis

Our algorithm has a natural method for model selection, the Q measure, which is the same objective function our embedding is based on. Since Normalized Cut is biased by the size of k, it cannot be used for choosing the best k which is one of the weaker areas of k-means algorithm.

It uses a greedy strategy of recursive bisection to search over different values of k. Because of this strategy it need not end as high quality a clustering (as high a Q value) as the other approach,but it will be faster since in going from k clusters tok+1only a portion of the data needs to be clustered rather than all of the data. This is explained in following steps:

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).

Case study

Once the leads are gathered from a suitable data collection algorithm also called as lead nurturing technique, we have the raw leads ready to be processed and distributed to advertisers. They can be processed manually or using data mining tools like WEKA – Waikato Environment for Knowledge Analysis. It is a machine learning open source software written in Java with user friendly visualization tools and algorithms for data analysis and predictive modeling. It is developed by machine learning group at University of Waikato, New Zealand (http://www.cs.waikato.ac.nz/ml/weka/bigdata.html).

Applications of data mining include prediction of the effectiveness of procedures, tests and result analysis and discovery of relationships among historical and current data to predict the trend of data flow/growth. These databases normally have huge amounts of information about user and their data/history/responses. Data mining techniques employed on these databases find relationships, helping the study of progression and providing predictive results. In this article, we will discuss a case study to show how Data Mining helps to classify and analysis of huge data with supervised learning techniques.

There are various steps involved in big data analysis starting from data collection, data cleansing, classification and up to pattern evaluation and trend report generation.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

There are many techniques for Big data analysis and there are wide spread algorithms and techniques practiced in Data mining world and one of the key technique is called Spectral Clustering. Spectral clustering is a graph theoretic technique for metric modification such that it gives much more global notion of similarity between data points as compared to other clustering methods such as k-means. It thus represents data in such a way that it is easier to find meaningful clusters on this new representation with inter-connected clusters.

We can understand this with an example of Facebook. In Facebook, we used to get suggested friends and suggested post which is typically based on clustered information from a user. This clustered information is based on location, age, school/college and friends link which is used to gather (mining related data) suggested friends or suggested posts depending on the predictive analysis. Also, the suggested post would be related area of interest like cricket, music etc., this kind of grouping related information in terms of clusters is called spectral clustering.

A typical example in an insurance based survey requires following data analytics points to be considered for data collection and refinement.

Function	Use Case Description	Metric
Managing Corporate Reputation	Social media analytics can be instrumental in managing the reputation of a business	Percentage increase in number of positive comments about your brand
	By real-time monitoring of customer posts and giving on time feedbacks, negative sentiments can be reduced	
Dynamic Profiling	Real-time monitoring of customer data will help in dynamic customer profiling and segmentation	ARPU(average revenue per user)
	This will enable business to target specific products at the customer groups thus increasing customer retention and satisfaction	
Precise Marketing	By leveraging customer data and feedbacks, business can sharpen their campaign management and pre-emptive churn avoidance mechanism effectively	Churn rate
Increased Upsell	Social network analysis by combining customer profiles, followers and their social circle activity can increase the upsell by targeting high-end products which can be afforded by the customers and their friends	ARPU(average revenue per user), MOU(minutes of use)
Pre-emptive Customer Care	Real-time analysis and monitoring of customer profiles in social media, business can track the changes in the lifestyle of the customers and provide effective solutions	Percentage reduction in discretionary expenses
Location based and personalized advertising	Geo-fencing advertising by analysing the location based information in social sites such as check-ins in Facebook and Foursquare	Net additions
Customized content services	Analysing customer feedbacks and surveys, personalized services such as tailored channel pack (set top box), recharge packs etc. can be developed	Increase in average revenue per user
Tariff Management	Analysis of subscribers profiles and surveys, telecos will be able to better manage their tariffs and be able to extract the maximum of cost-value trade off by the customers	Increase in average revenue per user
Identifying Key Customers	Tracking the number of followers for subscribers key customers who act as influencers can be identified	Increase in customer retention rate and revenue
	This information can be leveraged by focusing promotion and other retention programs on the identified key customers	

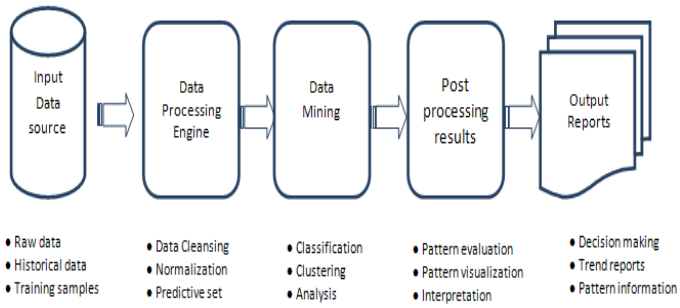
Result of evaluation

According to Vladimir Estivill-Castro, the notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms.[4] There is a common denominator: a group of data objects. However, different researchers employ different cluster models, and for each of these cluster models again different algorithms can be given. The notion of a cluster, as found by different algorithms, varies significantly in its properties. Understanding these "cluster models" is key to understanding the differences between the various algorithms. The key idea in Lead generation using a suitable data mining technique is to prepare required output based on raw data as there are difficulties in interpreting raw data such as

1) Data is stored in different sources in distributed locations.
2) Users find difficulty in locating the reports needed by them.
3) User interface for the current operational system is not satisfactory and is confusing and hard to use for decision makers.
4) It is very difficult, when the consolidated report from two or more different subject area is required.
5) There is no easy way to get assistance

Post processing results

Once we did our evaluation of data using Data mining techniques using M-clustering implementation in weka, we will have classified and clustered information in a more predictive set of results. This will be used further in post processing the results such as pattern evaluation (relationship and matching terms in the result cluster eg: average income for customers showing interest in insurance) and pattern visualization (trend reports).



Let us imagine a retail store located in a certain vicinity. Data Analyzer wish to understand the demographic profile and shopping preferences of the customers, who visit the store or those of prospects who live in the store's locality. This understanding will influence a whole range of business parameters in the store, such as the kind of products to stock, the nature of the promotions to make and even the amount of parking space to create.

In order to gain this understanding the user has to commission a detailed survey in which, say a 1000 respondents, provide answers to a whole range off questions ranging from their ethnicity (Demographic parameter) to their favorite means of transport to the store (behavioral parameter).

In such case study of retail domain of data mining and marketing analysis, one may infer following business improvement ideas for the marketing team to ideate the trends and improvise the market reach to customer.

- Quickly identify potential customers.
- Data mining the customer data for insights that drive new strategies for customer acquisition, retention, campaign optimization and next best offers.
- Send tailored recommendations to mobile devices at just the right time, while customers are in the right location to take advantage of offers.

- Generate discount coupons at the point of sale based on the customer's current and past orders and ensuring a higher redemption rate.
- Recalculate entire risk portfolios on the fly and understand future possibilities to mitigate risk.
- Analyze data from social media to detect new market trends and changes on demand.
- Use click stream analysis and data mining to detect fraudulent behaviour.

Conclusion

Data mining involves in clustered information gathered as 'raw data' from customers from various forums like social networking, trends in browsing pages, trends in search or pages visited. Analyzing such raw clusters of data which is huge in volume not only involves various analytics algorithm or techniques but also involves in filtering various required preference set which makes the base of data analysis. Clustering algorithms helps in such a condition where we focus on our analysis area and collect required subset of volumes of data gathered from Lead generation and process them to filter the preference set and produce the required results in terms of reports, diagrams, trend analysis and statistical data points.

References

[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 RobertJenssen, TorbjørnEltoft 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] GuoWensheng 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]M. Newman and M. Girvan. Finding and evaluatingcommunity structure in networks.Physical Review E,69, 026113 (2004).

[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]PengJin ,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] ZhijieXu, Laisheng Wang, JianchengLuo 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.