Abstract

In order to create a “user friendly” interface for organization of massive multidimensional data we propose a modified clustering mechanism which produces reliable clusters of very high quality. Graphical environment provides extreme comfort for users, immediate visualization and browsing of expected results on the screen. We created synthetic and real datasets for which we produce a variety of visualization clustering such as random clustering, structure-based clustering (using edges only), attribute-based clustering (using attributes only), combined clustering.

Introduction

Clustering is a common methodology for analysis of multidimensional data in many fields including genome biology. There are no clustering methods universally applicable for the variety of structures in multidimensional data sets, but it allows to discover meaningful structures in data in other words to create clusters. Main goal is to organize data into clusters which are built by objects with maximal degree of association if they belong to the same cluster and minimal otherwise. It should be noted that cluster analysis simply discovers structures without explaining why they exist.

Motivation and Implementation of the Use of Prefuse

To support visual representation group of related items in a large collection of data and to build user interactive systems for modeling the clusters, it was necessary to create a program that builds a dynamic force-directed graph layout. There were strict requirements for choosing the correct software or package:

- open source;
- force-directed layout;
- spring parameters can be set on a per-edge or per-edge-type basis;
- change visual attributes of edges and nodes (including creating "invisible" nodes and edges);
- manually move/pin nodes in graph;
- associate attributes with objects/nodes;
- integrate easily with other algorithms and software (e.g., clustering);
- java;
- well documented;
- scalable to several hundred nodes;

Short Analysis: Package That Allowed Building a Force-Directed Graph Layout.

On the internet, there were different visualization systems [1], [5], [6], which are thoroughly analyzed and compared in order to be ready for use, such as Graphiz [7], JGraph, Grappa, TouchGraph, JUNG [1], Prefuse [8], etc.

Graphiz [7] is a Java package with full Java graph data structures. It allows the user to draw a very aesthetic and really nice graphs and clusters. But, unfortunately, it does not allow drawing a dynamic layout. Grappa has serious problems with Java applet performance when graphs have hundreds of objects.
The same problem occurs with TouchGraph.
The JUNG seems useful, but its libraries don't include codes for clustering.
Based on our analysis the Prefuse seems to be the most reasonable package for the implementation of the project.

A short description of the Prefuse package:

Prefuse toolkit is a user interface toolkit for crafting interactive visualization applications of structured and unstructured data. Prefuse architecture allows you to create animated graphical interfaces for visualizing, exploring, and manipulating various forms of data; offers force simulation, graphics transformation, including panning and zooming. The Prefuse extensible library supports writing new layouts by composing existing modules, force-directed layout, integration of color maps for assigning colors to data elements, components interaction for common interaction including drag controls.
It also includes an extensible library for force-based simulation (nodes exert anti-gravity, edges act as springs).

Prefuse is open-source software. It is written in Java programming language using the Java2D graphics library and is designed to integrate with any application written using the Java Swing user interface library.

Prefuse allows the creation of "fast-drawing" and well designed force-directed layouts with several hundred nodes, clustering and filtering data and meets all of the requirements plus so much more.

Prefuse implementation

Based on Prefuse I created a package edu.umbc.cs.maple.graphlayout_N that allows the programming of a force-directed layout to produce a variety of target clustering:

- Random clustering
- Structure-based clustering (using edges only)
- Attribute-based clustering (using attributes only)
- Combiner clustering (using both)

This package was under development during the whole internship. The first draft of the program is available for viewing at my website – http://maple.cs.umbc.org/~nlozova, under the “project” tab under “Project Application in Java” (Figures 1-3). Further research and development of the project required additional visualization support and the ability to be more adaptable to the needs of others clusters creating and cluster analyzing.

A much more complex graph layout is posted on my website under the “Project” tab. The results are represented in Figures 4-20. Detailed description of the package I created for building such application is located on my website under the project tab, under “Package Descriptions”; filename: edu.umbc.cs.maple.graphlayout_N.doc

Clustering techniques

Modified k-means

According to the literature [11], [17], [21] the commonly used k-mean algorithm gives correct results for one, two or three dimensional data. This technique can be optimized and enhanced for typical applications by firstly applying to the synthetic data sets with further application to the real data sets. Based on our analysis it can be concluded that this algorithm will fail for the sets of higher dimensionality. The clear confirmation can be found in [30].
To overcome this problem we created a modified k-means algorithm with randomly chosen centroids of clusters according to the Gaussian distribution surrounded by n-dimensional nodes distributed by Gaussian distribution which belongs to the same cluster. As a result meaningful, high quality clusters are created without overlapping.

Structured-based clustering

Often the connection between nodes in the graph must be highlighted. Two simple techniques were used for creating structure based clustering: hovered node becomes the center of the cluster which consists of nodes which connected to that centroid. First of our techniques allows to pick out the connection and nodes using simple visualization (coloring) while hover over in the graph layout (see Fig.3). Other technique uses clicking on the certain node which moves it to the center of the screen. This node becomes the centroid of edge connected cluster surrounded by connected nodes. The nodes which do not belong to this cluster spread away.

Other techniques

Our force directed layout which uses coloring of nodes and images allows us to implement hybrid data mining method by simple adding of well known clustering algorithms.

Cluster’s Quality

Clustering appears to be a commonly used approach if the problem for analysis of data arrays should arise [2-4], [11-16], [21], [24], [27]. Depending on which clustering algorithm is chosen, one can create more or less accurate and stable clusters [12]. This is

called a cluster quality which varies from experiment to experiment and contains of cluster accuracy and cluster stability. These are used to assess performance of different clustering algorithms.

Back to 1971, Rand proposed [28] so called Rand index (which lies between 0 and 1) as measure of cluster accuracy. It shows how an agreement is related to total value of agreement plus disagreement. In other words, it defines what part of total number of pairs of objects is related to the pairs of objects which are either in the same groups in both partitions or in different groups in both partitions. Rand index reaches its maximum value of 1 if a perfect agreement appears.

A problem of Rand index of the two random partitions is that expected value does not take a constant (for example zero). This problem was solved by Hubert and Arabie [10] by creating a so called adjusted Rand index. It takes the zero if the index equals its expected value. Ka Yee Yeung et al. [30] show that adjusted Random index typically is much lower than Rand index. Hence range of possible values for adjusted index, which includes also zero, is wider it allows the calculation to reach higher sensitivity comparing to the Rand index. Accordingly to the commonly supported point of view, adjusted Rand index is the best assessment of cluster quality and I adopt it for our research as well.

I created a Java package (see Packages description on my web site edu.umbc.cs.maple.adj_rand_index_N) which will be used for our future data analysis at repeated experiments with different clustering algorithms.

Future Work

Our project is still under development by MAPLELAB. We suggest adding the adjusted Rand index into clustering analysis and doing multiple measurements for the same data set created. The results are supposed to be presented on 2006 International Conference on Intelligent User Interfaces, Sydney, Australia.

Future investigations will be extended for large number of nodes. These layouts might be dispersed using separate windows for each created cluster. The data of each cluster may be saved into an output file which can be used as an input file for further investigation.

I propose to extend our project and include such tasks as:
a) The main layout can be spit into different layouts as separate windows for each cluster;
b) Big clusters can include a lot of sub clusters. We can develop an algorithm to create sub clusters for further analyzing.

Acknowledgements

First of all, I would like to thank the Computer Research Association's Committee on the Status of Women in Computing Research (CRA-W) for sponsoring the Distributed Mentor Project (DMP).

Thanks to my teachers – Prof. D. Kraft and my math instructor, Martin Forrest, both of whish trusted my research abilities and recommended me for the DMP internships.

Thanks to my mentor at University of Maryland, Baltimore County (UMBC), Marie desJardins, who was always full of ideas, never allowed to relax. My project turned out the way it is because of her constant persuasion and would be where at the status it currently has; I think the project is amazing.

Thanks to the PhD students who greatly helped me out in critical points of the project at the Maple Lab at UMBC– Blazej Bulka and Qianjun Xu.

Thanks to my husband who found the opportunity for this internship; and my son – who helped me on some major points of creating the website.

Finally, thanks to Jeffery Heer, author of the wonderful prefuse toolkit package.

References

[1] www.aiSee.com – Graph Visualization. "*A picture is worth a thousand words.*" www.aiSee.com

[2] Basu, Sugato. Bilenko, Mikhail. Mooney, Raymond J. "*A Probabilistic Framework for Semi-Supervised Clustering.*" http://www.cs.unc.edu/Courses/comp290-90-f04/papers/p59-basu.pdf

[3] Cohn, David. Caruana, Rich. "*Semi-supervised Clustering: Incorporating User Feedback to Improve Cluster Utility.*"

[4] Eisen, Michael. "*Cluster 3.0 Manual.*" http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/cluster3.pdf

[5] Frasincar, Flavius. Telea, Alexandru. Houben, Geert-Jan. "*Adapting graph visualization techniques for the visualization of RDF data.*" http://wwwis.win.tue.nl/~hera/papers/VISSW2005/vsw2005.pdf

[6] Gansner, Emden R. North, Stephen C. "*An open graph visualization system and its applications to software engineering.*" http://www.graphviz.org/Documentation/GN99.pdf

[7] Graphviz.org. "*Graphviz - Graph Visualization Software.*" http://www.graphviz.org/About.php

[8] Heer, Jeffrey. Card, Stuart K. Landay, James A. "*Prefuse: a toolkit for interactive information visualization.*" http://guir.berkeley.edu/pubs/chi2005/prefuse.pdf

[9] Hotho, Andreas. Staab, Steffen. Stumme, Gerd. "*Explaining Text Clustering Results using Semantic Structures*." http://www.kde.cs.uni-kassel.de/hotho/pub/HothoStaabStumme_ExplainingTextClustering.pdf

[10] Hubert L and Arabie P., "*Comparing partitions*". Journal of Classification, 1985, 193-218.

[11] Jain, A.K. Murty, M.N. Flynn, P.J. "*Data Clustering: A Review*." http://scgwiki.iam.unibe.ch:8080/SCG/uploads/596/p264-jain.pdf

[12] Ka Yee Yeung, Mario Medvedovic and Roger E Bumgarner, "*Clustering gene-expression data with repeated measurements*", http://genomebiology.com/2003/4/5/R34

[13] Monti, Stefano. Tamayo, Pablo. Mesirov, Jill. Golub, Todd. "*Consensus Clustering: A resembling-based method for class discovery and visualization of gene expression microarray data.*" http://www.cs.utexas.edu/users/ml/biodm/papers/MLJ-biodm2.pdf

[14] Neville, Jennifer. Alder, Micah. Jensen, David. "*Clustering Relational Data Using Attribute and Link Information.*" http://kdl.cs.umass.edu/papers/neville-et-al-textlink2003.pdf

[15] Neville, Jennifer. Alder, Micah. Jensen, David. "*Spectral Clustering with Links and Attributes.*" http://kdl.cs.umass.edu/papers/neville-et-al-tr0442.pdf

[16] Pavagada, Ravi. Purvee, Edwin. Nanda, Amit. "*Clustering based on Semantic Relationships in Graph Visualization.*" http://webster.cs.uga.edu/~nanda/8380/Present.ppt

[17] Pelleg, Dan. Moore, Andrew. "*Accelerating Exact k-means Algorithms with Geometric Reasoning.*" http://delivery.acm.org/10.1145/320000/312248/p277-pelleg.pdf?key1=312248&key2=1891143211&coll=GUIDE&dl=GUIDE&CFID=50213129&CFTOKEN=90119830

[18] Pllu, Martin. "*Developing open source Java applications with java.net and Eclipse.*" https://eclipse-tutorial.dev.java.net/eclipse-tutorial/part1.html

[19] Rome, Jayson. "*Data Mining = Data Description: Requirements for a Large Scale Data Mining Application.*" http://prl.cs.gc.cuny.edu/web/LabWebsite/Rome/jrome/Slides_Tutorials/SyllogyProject3_5.pdf

[20] StatSoft, Inc. "*Basic Statistics.*" http://www.statsoft.com/textbook/stbasic.html

[21] StatSoft, Inc. "*Cluster Analysis.*" http://www.statsoft.com/textbook/stcluan.html

[22] StatSoft, Inc. "*Data Mining Techniques*."
http://www.statsoft.com/textbook/stdatmin.html

[23] Storkel, Scott. "*An Introduction to the Eclipse IDE*."
http://www.onjava.com/pub/a/onjava/2002/12/11/eclipse.html

[24] Stuart, Josh. "*Clustering and Cluster Evaluation*."
http://www.soe.ucsc.edu/classes/bme210/Winter04/lectures/Bio210w04-Lect10-Classifying.pdf

[25] Vrajitoru, Dana. DeBoni, Jason. "*Consistent Graph Layout for Weighted Graphs.*"
http://www.cs.iusb.edu/~danav/papers/172_Vrajitoru.pdf

[26] Wadstaff, Kiri. Cardie, Claire. "*Clustering with Instance-level Constraints.*"
http://www.litech.org/~wkiri/Papers/wagstaff-constraints-00.pdf

[27] Wadstaff, Kiri. Cardie, Claire. Rogers, Seth. Schroedl, Stefan. "*Clustering K-means Clustering with Background Knowledge*." http://www.litech.org/~wkiri/Papers/wagstaff-kmeans-01.pdf

[28] W.M. Rand, "*Objective criteria for the evaluation of clustering methods*", Journal of the American Statistical association, **66**, 846-850

[29] Yeung, Ka Yee. Medvedovic, Mario. Bumgarner, Roger E. "*Clustering gene-expression data with repeated measurements.*"
http://expression.microslu.washington.edu/expression/kayee/cluster2003/results_kayee3.pdf

[30] Yeung, Ka Yee. Ruzzo, Walter L. "*Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper 'An empirical study on Principal Component Analysis for clustering gene expression data*.' (to appear in Bioinformatics)."
http://faculty.washington.edu/kayee/pca/supp.pdf