ISBN 978-93-5156-328-0
International Conference of Advance Research and Innovation (ICARI-2014)

A Study on Feature Selection and Extraction

Surender Kumar

[a] Department of Computer Science & Engineering, Shri Venkateshwara University, Gajraula, India

Article Info

Article history:
Received 29 December 2013
Received in revised form
10 January 2014
Accepted 20 January 2014
Available online 1 February 2014

Keywords

Extractions,
Segmentation,
Image Segmentation,
Coloring,
Color layout,
High Dimensional Indexing

Abstract

This paper describes the feature selection and extraction mining functions. Oracle Data Mining supports a supervised form of feature selection and an unsupervised form of feature extraction.

1. Introduction

Sometimes too much information can reduce the effectiveness of data mining. Some of the columns of data attributes assembled for building and testing a model may not contribute meaningful information to the model. Some may actually detract from the quality and accuracy of the model.

For example, you might collect a great deal of data about a given population because you want to predict the likelihood of a certain illness within this group. Some of this information, perhaps much of it, will have little or no effect on susceptibility to the illness. Attributes such as the number of cars per household will probably have no effect whatsoever.

Irrelevant attributes simply add noise to the data and affect model accuracy. Noise increases the size of the model and the time and system resources needed for model building and scoring.

Moreover, data sets with many attributes may contain groups of attributes that are correlated. These attributes may actually be measuring the same underlying feature.

Their presence together in the build data can skew the logic of the algorithm and affect the accuracy of the model.

Wide data (many attributes) generally presents processing challenges for data mining algorithms. Model attributes are the dimensions of the processing space used by the algorithm. The higher the dimensionality of the processing space, the higher the computation cost involved in algorithmic processing

To minimize the effects of noise, correlation, and high dimensionality, some form of dimension reduction is sometimes a desirable preprocessing step for data mining. Feature selection and extraction are two approaches to dimension reduction.

- Feature selection — Selecting the most relevant attributes
- Feature extraction — Combining attributes into a new reduced set of features

2. Feature Selection

Oracle Data Mining supports feature selection in the attribute importance mining function. Attribute importance is a supervised function that ranks attributes according to their significance in predicting a target.

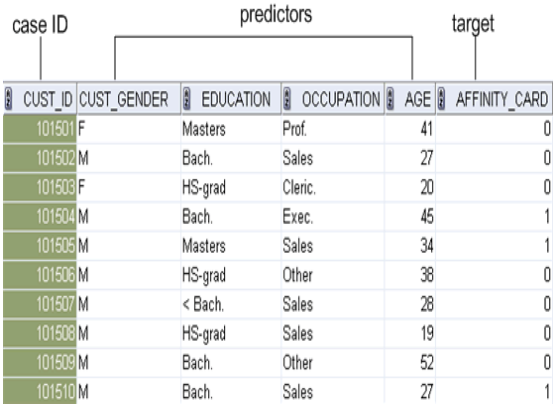
Finding the most significant predictors is the goal of some data mining projects. For example, a model might seek to find the principal characteristics of clients who pose a high credit risk.

Attribute importance is also useful as a preprocessing step in classification modeling, especially for models that use Naive Bayes or Support Vector Machine. The Decision Tree algorithm includes components that rank attributes as part of the model build.

*** Corresponding Author,**
E-mail address: surenderjaglan87@gmail.com

3. Data for Attribute Importance

Figure shows six columns and ten rows from the case table used to build the Oracle Data Mining sample attribute importance model, ai_sh_sample



CUST_ID	CUST_GENDER	EDUCATION	OCCUPATION	AGE	AFFINITY_CARD
101501	F	Masters	Prof.	41	0
101502	M	Bach.	Sales	27	0
101503	F	HS-grad	Cleric.	20	0
101504	M	Bach.	Exec.	45	1
101505	M	Masters	Sales	34	1
101506	M	HS-grad	Other	38	0
101507	M	< Bach.	Sales	28	0
101508	M	HS-grad	Sales	19	0
101509	M	Bach.	Other	52	0
101510	M	Bach.	Sales	27	1

Fig: 1 Sample Build Data for Attribute Importance

3.1 Color

The color feature is one of the most widely used visual features in image retrieval. It is relatively robust to background complication and independent of image size and orientation.

In image retrieval, the color histogram is the most commonly used color feature representation. Statistically, it denotes the joint probability of the intensities of the three color channels. Swain and Ballard proposed histogram intersection, an L 1 metric, as the similarity measure for the color histogram. To take into account the similarities between similar but not identical colors, Ioka and Niblack et al. introduced an L 2 -related metric in comparing the histograms. Furthermore, considering that most color histograms are very sparse and thus sensitive to noise, Stricker and Orengo proposed using the cumulated color histogram. Their research results demonstrated the advantages of the proposed approach over the conventional color histogram approach.

4. Feature Extraction

Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes.

The feature extraction process results in a much smaller and richer set of attributes. The maximum number of features is controlled by the FEAT_NUM_FEATURES build setting for feature extraction models.

Models built on extracted features may be of higher quality, because the data is described by fewer, more meaningful attributes.

Feature extraction projects a data set with higher dimensionality onto a smaller number of dimensions. As such it is useful for data visualization, since a complex data set can be effectively visualized when it is reduced to two or three dimensions.

Some applications of feature extraction are latent semantic analysis, data compression, data decomposition and projection, and pattern recognition. Feature extraction can also be used to enhance the speed and effectiveness of supervised learning.

Feature extraction can be used to extract the themes of a document collection, where documents are represented by a set of key words and their frequencies. Each theme (feature) is represented by a combination of keywords. The documents in the collection can then be expressed in terms of the discovered themes.

4.1 Color Layout

Although the global color feature is simple to calculate and can provide reasonable discriminating power in image retrieval, it tends to give too many false positives when the image collection is large. Many research results suggested that using color layout (both color feature and spatial relations) is a better solution to image retrieval. To extend the global color feature to a local one, a natural approach is to divide the whole image into sub blocks and extract color features from each of the sub blocks.

A variation of this approach is the quadtree based color layout approach, where the entire image was split into a quadtree structure and each tree branch had its own histogram to describe its color content. Although conceptually simple, this regular sub block-based approach cannot provide accurate local color information and is computation and storage-expensive.

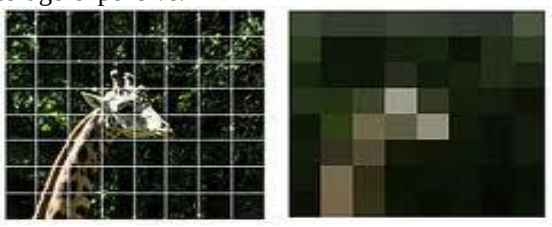


Fig: 2. Color layout for 300Px Image

5. Segmentation

Segmentation is very important to image retrieval. Both the shape feature and the layout

feature depend on good segmentation. In this subsection we will describe some existing segmentation techniques used in both computer vision and image retrieval.

In Lybanon et al. researched a morphological operation (opening and closing) approach in image segmentation. They tested their approach in various types of images, including optical astronomical images, infrared ocean images, and magneto grams. While this approach was effective in dealing with the above scientific image types, its performance needs to be further evaluated for more complex natural scene images.

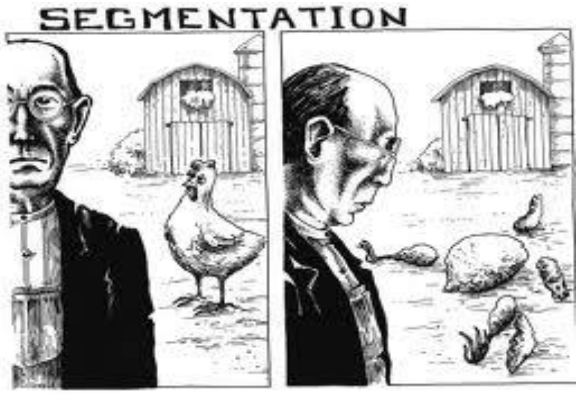


Fig: 3. Segmentation of an Image

In Hansen and Higgins exploited the individual strengths of watershed analysis and relaxation labeling. Since fast algorithm exists for the watershed method, they first used the watershed to subdivide an image into catchmen basins. They then used relaxation labeling to refine and update the classification of catchment basins initially obtained from the watershed to take advantage of the relaxation labeling's robustness to noise.

This approach is based on the fact that local entropy maxima correspond to the uncertainties among various regions in the image. This approach was very effective for images whose histograms do not have clear peaks and valleys. Other segmentation techniques based on Delaunay triangulation, fractals, and edge flow.

6. High Dimensional Indexing

To make the content-based image retrieval truly scalable to large size image collections, efficient multidimensional indexing techniques need to be explored. There are two main challenges in such an exploration for image retrieval. High dimensionality the dimensionality of the feature vectors is normally of the order of 10Non-Euclidean similarity measure. Since Euclidean measure may not effectively simulate human perception of a certain visual content, various other similarity measures, such as histogram intersection, cosine, correlation, need to be supported.

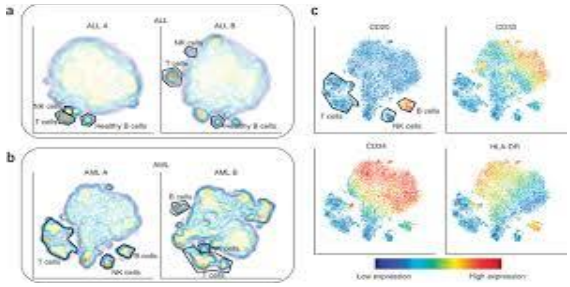


Fig: 4. High Dimensional Indexing for Cancer samples

Towards solving these problems, one promising approach is to first perform dimension reduction and then to use appropriate multidimensional indexing techniques, which are capable of supporting non-Euclidean similarity measures.

References

[1] MPEG-7 applications document, ISO/IEC JTC1/SC29/WG11 N1922, MPEG97, 1997
[2] MPEG-7 context and objectives, ISO/IEC JTC1/SC29/WG11 N1920, MPEG97, (5), 1997
[3] Retrievalware, demo page http://vrw/excalib.com /cgibin/sdk/cst/cst2.bat, 1997
[4] Special issue on visual information management, R. Jain, Guest Ed., Comm. ACM, 1997
[5] Third draft of MPEG-7 requirements, ISO/IEC JTC1/SC29/WG11 N1921, MPEG97, 1997
[6] Proc. Int. Conf. on Multimedia, ACM, New York, 1997
[7] A. D. Alexandrov, W. Y. Ma, A. El Abbadi, B. S. Manjunath, Adaptive filtering and indexing for iamge databases, in Proc. SPIE Storage and Retrieval for Image and Video Databases, 1995
[8] J. Allan, Relevance feedback with too much data, in Proc. of SIGIR'95, 1995
[9] E. M. Arkin, L. Chew, D. Huttenlocher, K. Kedem, J. Mitchell, An efficiently computable metric for comparing polygonal shapes, IEEE Trans. Patt. Recog. Intell. 13(3), 1991
[10] V. Athitsos, M. J. Swain, C. Frankel, Distinguishing, photographs and graphics on the world wide web, in Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries, 1997
[11] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, C. F. Shu, The Virage image search engine: An open framework for image management, in Proc. SPIE Storage and Retrieval for Image and Video Databases
[12] H. G. Barrow, Parametric correspondence and chamfer matching: Two new teachniques for

image matching, in Proc. 5th Int. Joint Conf. Artificial Intelligence, 1977

[13] [13]. N. Beckmann, H.-P. Kriegel, R. Schneider, B. Seeger, The R∗ -tree: An efficient and robust access method for points and rectangles, in Proc. ACM SIGMOD, 1990

[14] M. Beigi, A. Benitez, S.-F. Chang, Metaseek: A content-based meta search engine for images, in Proc SPIE Storage and Retrieval for Image and Video Databases, San Jose, CA, 1998, demo and document: http://www.ctr.columbia.edu/metaseek

[15] G. Borgefors, Hierarchical chamfer matching: A parametric edge matching algorithm. IEEE Trans. Patt.Recog. Intell., 1988

[16] C. Buckley, G. Salton, Optimization of relevance feedback weights, in Proc. of SIGIR'95, 1995

[17] J. P. Callan, W. B. Croft, S. M. Harding, The inquery retrieval system, in Proc. of 3rd Int. Conf. on Database and Expert System Application, 1992

[18] C. Carson, S. Belongie, H. Greenspan, J. Malik, Region-based image querying, in Proc. of IEEE Workshop on Content-Based Access of Image and Video Libraries, in Conjuction with IEEE CVPR'97, 1997

[19] S. Chandrasekaran, B. S. Manjunath, Y. F. Wang, J. Winkeler, H. Zhang, An eigenspace update algorithm for image analysis, CVGIP: Graphical Models and Image Processing Journal, 1997

[20] N. S. Chang and K. S. Fu, A Relational Database System for Images, Technical Report TR-EE 79-28, Purdue University, 1979.

[21] N. S. Chang, K. S. Fu, Query-by pictorial-example, IEEE Trans. on Software Engineering SE-6(6), 1980

[22] S.-F. Chang, A. Eleftheriadis, R. McClintock, Next-generation content representation, creation and searching for new media applications in education, IEEE Proceedings, 1998, to appear

[23] S.-F. Chang, J. R. Smith, M. Beigi, A. Benitez, Visual information retrieval from large distributed online repositories. Comm. ACM (Special Issue on Visual Information Retrieval) 1997, 12–20.

[24] S. K. Chang, Pictorial data-base systems, IEEE Computer, 1981

[25] S. K. Chang, A. Hsu, Image information systems: Where do we go from here? IEEE Trans. on Knowledge and Data Engineering 4(5), 1992

[26] S. K. Chang, C. W. Yan, D. C. Dimitroff, T. Arndt, An intelligent image database system, IEEE Trans Software Eng. 14(5), 1988

[27] S. F. Chang, Compressed-domain content-based image and video retrieval, in Proc. Symposium on Multi- media Communications and Video Coding, 1995

[28] S. F. Chang, Compressed-domain techniques for image/video indexing and manipulation, in Proc. ICIP95 Special Session on Digital Library and Video on Demand, 1995

[29] S.-F. Chang J. R. Smith, Finding images/video in large archives D-Lib. Magazine, 1997

[30] S. F. Chang, J. Smith, Extracting multidimensional signal features for content-based visual query, in Proc. SPIE Symposium on Visual Communications and Signal Processing, 1995

[31] T. Chang and C.-C. J. Kuo, Texture analysis and classification with tree-structured wavelet transform, IEEE Trans. Image Proc. 2(4), 429–441, 1993