International Journal of Knowledge Engineering
ISSN: 0976-5816 & E-ISSN: 0976-5824, Volume 3, Issue 2, 2012, pp.-188-192.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000230

KNOWLEDGE DISCOVERY PROCESS IN THE IMAGE-SEGMENTATION DATA

SRIMANI P.K.[1] AND SHANTHI M.[2]*
[1]Dept. of Computer Science & Maths, Bangalore University, Bangalore-560 078, Karnataka, India.
[2]Dept. of Information Science & Engineering, Atria Institute of Technology, Bangalore-560 024, Karnataka, India.
*Corresponding Author: Email- shanthi_md@yahoo.co.in

Received: October 25, 2012; Accepted: November 06, 2012

Abstract- This paper discusses in detail the behavior of the different classification on image segmentation data. The result predicts the different aspects of the classification model. It is found that NNEG is the best classifier with accuracy of 96.2771%. $ROC|_{max}$ and $ROC|_{min}$ are computed for different classes and are found to be interesting.

Keywords- Data mining, knowledge discovery, classifier, accuracy, confusion matrix, image segmentation.

Citation: Srimani P.K. and Shanthi M. (2012) Knowledge Discovery Process in The Image-Segmentation Data. International Journal of Knowledge Engineering, ISSN: 0976-5816 & E-ISSN: 0976-5824, Volume 3, Issue 2, pp.-188-192.

Copyright: Copyright©2012 Srimani P.K. and Shanthi M. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

Introduction

Data mining is a knowledge discovery process in databases [4] or KDD, which is a relatively young, popular and interdisciplinary field of computer science [1,2]. Further, it is the process of discovering new patterns from large data sets involving methods at the intersection of machine learning, artificial intelligences neural networks, databases systems and computational statistic [2]. In fact, data mining aims at extracting knowledge from a data set in a human-understandable structure [1] and involves management of database and data, data preprocessing, model, inference and complexity considerations, post-processing of resulting structure, visualization and online updation.

Related work: Very sparse literature pertaining to the subject of research is available. Some of the works include [3,8-11].

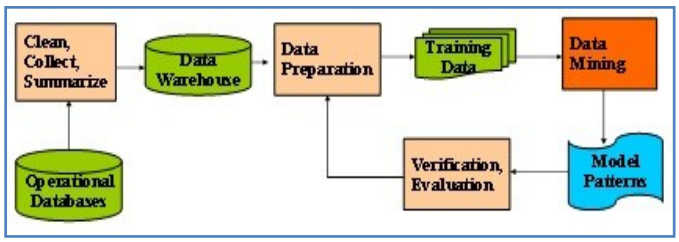

Fig. 1- Steps in KDD

KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of information from data that are previously unknown implicit and potentially useful. While data mining and KDD are often treated as equivalent words but in real data mining is an important step in the KDD process. The [Fig-1] shows data mining as a step in an iterative knowledge discovery process.

The knowledge discovery in databases (KDD) process is commonly defined with the stages

i) Selection, ii) Preprocessing iii) Transformation iv) *Data Mining* v) Interpretation/Evaluation

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge [2]. The steps involved in the iterative process are the following:

1. **Data Cleaning:** also known as data cleansing is a phase in which noise data and irrelevant data are removed from the collection.
2. **Data Integration:** is the step which combines multiple data sources of heterogeneous as a common source.
3. **Data Selection:** is this step in which the data relevant to the analysis is decided on and retrieved from the data collection.
4. **Data Transformation or Data Consolidation:** is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
5. **Data Mining:** is the crucial step in which clever techniques are applied to extract patterns potentially useful.
6. **Pattern Evaluation:** is the step in which, strictly interesting patterns representing knowledge are identified based on given measures.

7. **Knowledge Representation:** is the final phase in which the discovered knowledge is visually represented to the user. In this step visualization techniques are used to help users understand and interpret the data mining results.

Data Mining Process

In the KDD process, the data mining methods are for extracting patterns from data. The patterns that can be discovered depend upon the data mining tasks applied. The two types of data mining tasks generally considered are: *descriptive data mining tasks* that describe the general properties of the existing data, and *predictive data mining tasks* that attempt to do predictions based on available data. Data mining is possible only on data sets that are of quantitative, textual, or multimedia nature.

Data Mining Tasks

The six common tasks involved in data mining are:

1. Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records and data errors which requires further investigation.
2. Association rule learning (Dependency modeling) – Searches for relationships between variables. This can be illustrated as follows: a supermarket might gather data on customer purchasing habits. It can determine by using association rule learning, what products are frequently bought together and the information could be use for marketing purposes. This is analysis is referred to as market basket analysis.
3. Clustering – is the task of discovering groups and structures in the data that are in some way or other "similar" (where the known structures are not used in the data).
4. Classification – is the task of generalizing the known structures for applying on new data. For example, an email might be classified as legitimate or spam by an email program.
5. Regression – Attempts to find a function which models the data with the least error.
6. Summarization – is the task of providing a more compact representation of the data set along with visualization and report generation for the input.

Different levels of analysis are available: (i) Artificial neural networks, (ii). Decision trees, (iii), Genetic algorithms (iv). Nearest neighbor method, (v). Rule induction, (vi). Data visualization.

Classification

Classification is the most important and frequently used technique in data mining. It facilitates the determination of models capable of describing and distinguishing classes or concepts. It is possible to represent the derived model in various forms such as classification (IF-THEN) rules, decision tree, artificial neural networking, etc.

In data mining, classification is one of the most important task. It is a supervised learning since it maps the data into predefined targets. The aim of performing classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Thus, the purpose of classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The methods that are commonly used for data mining classification tasks are found in groups [5].

A decision tree is a flowchart like tree structure where each node denotes a test on an attribute value, each branch represents an outcome of the test and tree leaves represent classes. In fact it is possible to convert decision trees into classification rules. Further, a neural network is typically a collection of neurons like processing units with weighted connections between the units when used for classification purpose. The system as to find the rules that predict the class from the prediction attributes during the learning process of classification rules. Therefore to start with the conditions for each class, the data mine system and the constructs descriptions for the classes are to be defined by the user. The system should infer rules that govern the classification once classes are defined.

Therefore the system should be able to find the description of the each class where, the description should only refer to the prediction attributes of the training set so that the positive examples should satisfy the description. If the description of the rule covers only the positive examples, then it is said to be correct.

In recent years, data mining has been used widely in the areas of science and engineering specifically medicine, education, bioinformatics, genetics and electrical power engineering. Top of Form

Methodology

Segmentation is the process of partitioning a *digital image* into multiple *segments* (superpixels). The purpose of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze [4]. For locating objects and boundaries in images, the process of image segmentation is used, which assigns a label to every pixel in an image such that pixels with the same label share certain visual characteristics.

The result of image segmentation is a set of segments that collectively cover the entire image, or a set of *contours* that are extracted from the image. It is observed that each of the pixels in a region have some similarities with respect to *color*, *intensity*, or *texture*. Certainly the adjacent regions have significantly different characteristic (s) [4]. In a *medical imaging*, image segmentation is applied to a stack of images and the resulting contours can be used to create 3D reconstructions with the help of interpolation algorithms like Marching cubes.

Several general-purpose algorithms and techniques have been developed for image segmentation and these techniques often have to be combined with domain knowledge in order to effectively solve an image segmentation problem for a problem domain since there is no general solution for the image segmentation problem.

Data Set Description

The Image-seg dataset

The information is a replica of the notes for the segmentation dataset from the UCI repository.

Title: Image Segmentation data Source Information

i. Creators: Vision Group, University of Massachusetts,
ii. Donor: Vision Group,

iii. Date: November, 1990

The instances were drawn randomly from a database of 7 outdoor images. Hand segmentation of the images where done to create a classification for every pixel. Each instance is a 3x3 region. No literature in this direction is available. The number of number of instances includes training data:210 and test data : 2100. There are 19 continuous attributes.

Attribute Information

1. region-centroid-col: the column of the center pixel of the region.
2. region-centroid-row: the row of the center pixel of the region.
3. region-pixel-count: the number of pixels in a region = 9.
4. short-line-density-5: the results of a line extraction algorithm that counts how many lines of length 5 (any orientation) with low contrast, less than or equal to 5, go through the region.
5. short-line-density-2: same as short-line-density-5 but counts lines of high contrast, above 5.
6. vedge-mean: measure the contrast of horizontally adjacent pixels in the region. This attribute is used as a vertical edge detector, which is used for horizontal line detection.
7. vegde-sd: (see 6)
8. hedge-mean: measures the contrast of vertically adjacent pixels. Used for horizontal line detection.
9. hedge-sd: (see 8).
10. intensity-mean: the average over the region of (R + G + B)/3
11. rawred-mean: the average over the region of the R value.
12. rawblue-mean: the average over the region of the B value.
13. rawgreen-mean: the average over the region of the G value.
14. exred-mean: measure the excess red: (2R - (G + B))
15. exblue-mean: measure the excess blue: (2B - (G + R))
16. exgreen-mean: measure the excess green: (2G - (R + B))
17. Value-mean: 3-d nonlinear transformation of RGB.
18. saturation-mean: (see 17)
19. hue-mean: (see 17)

Classes: brickface, sky, foliage, cement, window, path, grass.

- 30 instances per class for training data.
- 300 instances per class for test data.
- There are no missing attribute values in class distribution.

Experiments and Results

This section discusses, the results of the experiments conducted on the image-seg dataset consisting of 2100 instances and 19 attributes are presented and discussed in detail.

1. The *data* and *test* files were combined and then stratified to ensure equal representation of the output classes in each of the Delve task-instance training sets.
2. Attribute 3 (region-pixel-count) was deleted since it is a constant for this dataset.

Conjunctive Rule: learns a single rule that predicts either a numeric or a nominal class value. The default class value of the uncovered training instances are assigned to the uncovered test instances. The computation of information gain (nominal class) or variance reduction (numeric class) is carried out and by using reduced error pruning, the rules are pruned.

[Tables-1a], [Tables-2a], [Tables-3a], [Tables-4a], [Tables-5a], [Tables-6a], [Tables-7a] predicts the TP rate, precision, F-measure and ROC area for different classes viz., brickface, sky, foliage, cement, window, path and grass. [Tables-1b], [Tables-2b], [Tables-3b], [Tables-4b], [Tables-5b], [Tables-6b], [Tables-7b] present the confusion matrices which predict the correctly and incorrectly classified instances.

In [Tables-1b], [Tables-2b], [Tables-3b], [Tables-4b], [Tables-5b], [Tables-6b], [Tables-7b] the notations used are: a, b, c, d, e, f, g <-- classified as a = brickface, b = sky, c = foliage, d = cement, e = window, f = path, g = grass

Table 1a- Conjunctive Rule

TP Rate	FP Rate	Precision	Recall	F-Measure	Roc Area	Class
0.7	0.394	0.228	0.7	0.344	0.737	brickface
0.8	0.252	0.346	0.8	0.484	0.836	Sky
0	0	0	0	0	0.701	Foliage
0	0	0	0	0	0.734	Cement
0	0	0	0	0	0.73	Window
0.2	0.035	0.489	0.2	0.284	0.862	Path
0.3	0.153	0.153	0.3	0.271	0.801	grass

Here $ROC|_{max}$ is for path and $ROC|_{min}$ is for foliage.

Table 1b- Confusion Matrix

a	b	c	d	e	f	g
231	0	0	0	0	0	99
66	264	0	0	0	0	0
223	17	0	0	0	0	90
92	218	0	0	0	4	16
230	3	0	0	0	0	97
4	260	0	0	0	66	0
166	0	0	0	0	65	99

JRip

JRip implements RIPPER by including the heuristic global optimization of the rule set.

Table 2a- JRip

TP Rate	FP Rate	Precision	Recall	F-Measure	Roc Area	Class
0.976	0.006	0.964	0.076	0.97	0.099	brickface
0.994	0	1	0.994	0.997	0.997	Sky
0.924	0.012	0.927	0.924	0.926	0.983	Foliage
0.933	0.007	0.96	0.933	0.946	0.978	Cement
0.897	0.016	0.902	0.897	0.9	0.975	Window
0.982	0.006	0.964	0.982	0.973	0.999	Path
0.994	0.003	0.982	0.994	0.988	0.998	grass

Here $ROC|_{max}$ is for path and $ROC|_{min}$ is for brickface.

Table 2b- Confusion matrix

a	b	c	d	e	f	g
322	0	1	4	3	0	0
1	328	0	0	0	1	0
3	0	305	1	17	4	0
4	0	4	308	9	2	3
4	0	17	7	296	4	2
0	0	2	1	2	324	1
0	0	0	0	1	1	328

PART

Part obtains rules from partial decision trees and builds the tree using C4.5's heuristics with the same user-defined parameters as J4.8.

Table 3a- PART

TP Rate	FP Rate	Precision	Recall	F-Measure	Roc Area	Class
0.976	0.08	0.955	0.976	0.966	0.988	brickface
1	0.001	0.997	1	0.998	1	Sky
0.939	0.012	0.928	0.939	0.934	0.964	Foliage
0.939	0.009	0.948	0.939	0.944	0.981	Cement
0.894	0.012	0.925	0.894	0.909	0.961	Window
0.997	0.001	0.994	0.997	0.995	0.999	Path
0.991	0.002	0.988	0.991	0.989	0.995	grass

Here $ROC|_{max}$ is for sky and $ROC|_{min}$ is for foliage.

Table 3b- Confusion matrix

a	b	c	d	e	f	g
322	0	2	4	2	0	0
0	330	0	0	0	0	0
2	1	310	3	12	0	2
5	0	6	310	9	0	0
7	0	16	10	295	1	1
0	0	0	0	0	329	1
1	0	0	0	1	1	327

OneR

OneR is classifier consisting of one parameter: the minimum bucket size for discretization.

Table 4a- OneR

TP Rate	FP Rate	Precision	Recall	F-Measure	Roc Area	Class
0.97	0.022	0.882	0.97	0.924	0.974	brickface
0.73	0.069	0.639	0.73	0.682	0.831	Sky
0.458	0.075	0.503	0.458	0.479	0.691	Foliage
0.391	0.08	0.449	0.391	0.418	0.656	Cement
0.409	0.081	0.456	0.409	0.431	0.664	Window
0.558	0.089	0.511	0.558	0.533	0.734	Path
0.991	0	1	0.991	0.995	0.995	grass

Here $ROC|_{max}$ is for grass and $ROC|_{min}$ is for cement.

Table 4b- Confusion matrix

a	b	c	d	e	f	g
320	0	0	3	4	3	0
0	241	63	13	11	2	0
5	105	151	11	53	5	0
23	2	19	129	53	104	0
13	8	51	61	135	62	0
0	21	15	70	40	184	0
2	0	1	0	0	0	327

Ridor

Ridor learns rules with exceptions by generating the default rule, using incremental reduced-error pruning to find exceptions with the smallest error rate, finding the best exceptions for each exception, and iterating.

It generates a default rule first and then the exceptions for the default rule with the least (weighted) error rate. Then it generates the "best" exceptions for each exception and iterates until it is error free. Thus a tree-like expansion of exceptions is performed where the exceptions are a set of rules that predict classes other than the default. Here IREP is used to generate the exceptions.

Table 5a- Ridor

TP Rate	FP Rate	Precision	Recall	F-Measure	Roc Area	Class
0.982	0.05	0.97	0.982	0.976	0.988	brickface
1	0.001	0.997	1	0.998	1	Sky
0.93	0.01	0.942	0.93	0.936	0.96	Foliage
0.912	0.007	0.959	0.912	0.935	0.951	Cement
0.921	0.02	0.886	0.921	0.903	0.951	Window
0.997	0.003	0.982	0.997	0.989	0.997	Path
0.988	0.001	0.997	0.988	0.992	0.994	grass

Here $ROC|_{max}$ is for sky and $ROC|_{min}$ is for cement and window.

Table 5b- Confusion matrix

a	b	c	d	e	f	g
324	0	0	0	6	0	0
0	330	0	0	0	0	0
2	1	307	6	14	0	0
4	0	3	301	19	2	1
4	0	16	6	304	0	0
0	0	0	1	0	329	0
0	0	0	0	0	4	326

NNGE

Nnge is a nearest-neighbor method for generating rules and uses non-nested generalized exemplars.

Table 6a- NNGE

TP Rate	FP Rate	Precision	Recall	F-Measure	Roc Area	Class
0.988	0.003	0.982	0.988	0.985	0.992	brickface
1	0	1	1	1	1	Sky
0.93	0.016	0.906	0.93	0.918	0.957	Foliage
0.955	0.01	0.943	0.955	0.949	0.972	Cement
0.87	0.013	0.917	0.87	0.893	0.928	Window
1	0	1	1	1	1	Path
0.997	0.002	0.991	0.997	0.994	0.998	grass

Here $ROC|_{max}$ is for path and $ROC|_{min}$ is for foliage.

Table 6b- Confusion matrix

a	b	c	d	e	f	g
326	0	0	2	2	0	0
0	330	0	0	0	0	0
3	0	307	3	17	0	0
1	0	6	315	6	0	2
2	0	26	14	287	0	1
0	0	0	0	0	330	0
0	0	0	0	1	0	329

Decision Table

Decision Table builds a decision table majority classifier and evaluates the feature subsets by using best-first search and use cross-validation for evaluation.

Table 7a- Decision Table

TP Rate	FP Rate	Precision	Recall	F-Measure	Roc Area	Class
0.967	0.01	0.944	0.967	0.955	0.998	brickface
0.997	0.001	0.994	0.997	0.995	1	Sky
0.879	0.017	0.898	0.879	0.888	0.989	Foliage
0.882	0.015	0.907	0.882	0.894	0.989	Cement
0.848	0.03	0.826	0.848	0.837	0.982	Window
0.991	0.002	0.991	0.991	0.991	1	Path
0.994	0	1	0.994	0.997	1	Grass

Here $ROC|_{max}$ is for path,sky and grass and $ROC|_{min}$ is for window.

Table 7b- Confusion matrix

a	b	c	d	e	f	g
319	0	0	7	4	0	0
0	329	0	1	0	0	0
7	1	290	4	27	1	0
3	1	5	291	28	2	0
7	0	26	17	280	0	0
1	0	1	1	0	327	0
1	0	1	0	0	0	328

Table 8- Final predictions

	CCI	ICI	KS	MAE	RMS	RRSE	Accuracy
Conjuctive Rule	660	1650	0.1667	0.2087	0.323	92%	28.57%
JRip	2211	99	0.95	0.016	0.107	31%	95.71%
PART	2223	87	0.9561	0.118	0.1	29%	96.23%
OneR	1487	823	0.5843	0.1018	0.319	91%	64.37%
Ridor	2221	89	0.9551	0.011	0.104	30%	96.15%
NNGE	2224	86	0.9566	0.0106	0.103	29%	96.28%
Decision Table	2164	146	0.9263	0.0272	0.118	34%	93.68%

In the above tables, the following abbreviations are used: CCI: Correctly Classified Instances; ICI: Incorrectly classified Instances; KS: Kappa Statistic; MAE: Mean Absolute error; RMSE: Root mean squared error; and RRSE: Root relative square error

[Table-8], presents in detail the error prediction, kappa statistics and accuracy predictions for all the seven classifiers viz., conjunctive rule, Jrip, PART, OneR, Ridor, NNGE and Decision table.

A glance at [Table-1] to [Table-8] reveals that the classifiers other than conjunctive rule and oneR, performs extremely well and the best classifier happens to be NNGE.

Conclusion

Image segmentation is the process of assigning a label to every pixel in an image such that the pixels with the same label share certain usual characteristics. For the present analysis, the instances were drawn randomly from a database of 7 outdoor images. The images were hand-segmented to create a classification for every pixel. The data set considered consists of 2100 instances and 19 continuous attributes. The classifiers considered in the experiment are: conjunctive Rule, PART, JRIP, NNGE, RIDOR, OneR and Decision Table. The results are presented in [Tables-1] to [Table-8] and finally it is concluded that NNGE is the best classifier.

Acknowledgement

One of us Mrs. Shanthi Mahesh is grateful to Bharathiar University, Tamil Nadu & Atria Institute of Technology for providing the facilities to carry out the research work.

References

[1] Clifton C. (2010) *Encyclopædia Britannica: Definition of Data Mining*, 12-09.

[2] ACM SIGKDD (2006) *Data Mining Curriculum*, 04-30.

[3] Diane J.C., Pullman N.K. and Selivn A. (2008) *PE & RS journal*.

[4] Fayyad U., Gregory P.S. and Padhraic S. (2008) *From Data Mining to Knowledge Discovery in Databases*, 12-17.

[5] Fukuda K. and Pearson A., *Environmental Science Programme*, University of Canterbury, New Zealand.

[6] James C.T., *Computational & Information Sciences and Technology Office* (606.3).

[7] Linda G.S. and George C.S. (2001) *Computer Vision*, 279-325, New Jersey.

[8] James C.T., *NASA Goddard Space Flight Center*, Greenbelt, MD 20771.

[9] Pham D.L., Xu Chenyang, Prince J.L. (2000) *Annual Review of Biomedical Engineering,* 2, 315-337.

[10] Soh Leen-Kiat and Tsatsoulis C. (1999) *CSE Journal,* 48.

[11] Wismiller A et al., *Neural Networks,* Neuroinformatik, Universit Bielefeld, Bielefeld, Germany.