

Anomaly Detection from Complex Data Sets

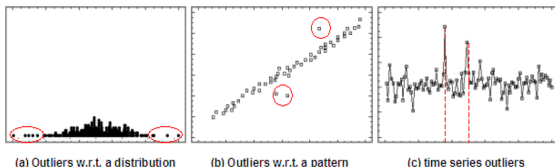
Khoi-Nguyen Tran (u4315673)

COMP3006 - Computer Science Research Project
Australian National University

Supervisor: Dr Huidong (Warren) Jin at NICTA

Introduction

- ▶ Anomaly: A data value that is irregular, inconsistent or discrepant relative to other values in a data set by some defined measure.



Introduction

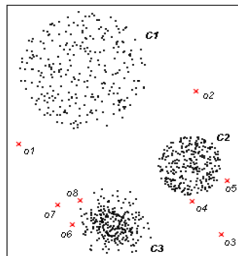
- ▶ Complex Data Sets: Structured or unstructured data with different types of attributes.

key	Price	broker	age	exterior	roof	arft	mbr	br1	br2	lr	dr	kt	str	pk	add	...
12345	\$95,000	Sussex	22	Stucco	Gmvel	911	13x9	13x8	0	14x12	12x9	9x7	Y	N
12346	\$110,000	Sutton	16	Mixed	Tar/Gr	939	13x10	13x9	6x5	11x13	12x11	9x5	Y	Y
12347	\$114,000	Rennie	10	Wood	Tar/Gr	933	11x13	10x10	0	12x13	12x9	10x7	N	Y
12348	\$119,900	Rennie	10	Wood	Tar/Gr	974	11x13	10x10	0	13x12	12x10	9x9	N	Y
12349	\$116,900	P.George	12	Stucco	Tar/Gr	901	12x12	11x10	8x3	15x12	11x9	9x7	Y	Y
12350	\$99,000	P.George	17	Stucco	Tar/Gr	879	13x10	12x9	0	13x11	10x10	6x11	Y	N
12351	\$119,500	Sutton	14	Mixed	Tar/Gr	815	14x11	14x9	0	13x12	7x9	9x7	N	Y
12352	\$115,000	HomeLife	6	Mixed	Tar/Gr	911	14x11	14x9	0	14x12	13x9	7x7	Y	Y
12353	\$116,900	Rennie	10	Wood/stc	Tar/Gr	964	11x13	14x9	0	14x11	12x9	9x7	N	Y
12354	\$110,500	Rennie	16	Mixed	Tar/Gr	990	13x11	13x8	0	12x13	10x10	17x5	N	Y
...

This project will focus on structured data with simple attribute types such as continuous, categorical and intervals.

Motivation

- ▶ Anomaly detection is an important part of data mining.
- ▶ Many real-world applications.
- ▶ Many difficulties arise from working with mixed-attribute data sets.
 - ▶ eg. Defining correlation between different types of attributes.



Plan

- ▶ Explore various ideas and implementations.
- ▶ Develop new ideas for the problem.
- ▶ Implement those ideas.
- ▶ Experiments to support the new ideas.

Reference

- ▶ Diagrams and table from lectures by Osmar Zaiane.
Available at
<http://www.cs.ualberta.ca/~zaiane/courses/cmp695/F07/>