Dependence Degree

and

Feature Selection

for Categorical Data

Wenxue Huang and Michael Vainder

Generation 5 Mathematical Technologies Inc.

wenxue@generation5.net,
michaelv@generation5.net

Motivation and Introduction

Traditionally the measure of association for cross-classification for categorical data takes a point view of variance or divergence. Goodman and Kruskal defined the association degree of Y with X as

$$(V(Y) - E(V(Y|X)))/V(Y).$$

We instead take an opposite angle of view: convergence or concentration.

This point of view has certain advantages. It allows us to see more directly and clearly how a variable is associated with another/others both locally (vertically) and globally (horizontally). From this point of view we introduce a new measure of association, referred to as *dependence degree*, and discuss one of its applications in data mining technologies: feature selection.

- General $m \to n$ association, in what degree (denoted by DepDeg) Y depends on X:

$n = 1$:	$\text{DepDeg}(Y\|X) = 1 = 100\%$;
$m = 1$:	$\text{DepDeg}(Y\|X) = E(p(Y))$;
General:	$\text{DepDeg}(Y\|X) =?$

- Nominal data is the most general among categorical data types

- Find a explanation base for the dependent variable Y for reducing dim.

- Find a structure base for a data set.

- Develop a dependent variable associated cluster analysis

- Develop a dependent variable associated dissimilarity measure for local approach

- Determine memory length of a stochastic process

1. Concepts

Definition. Let x and y be two categorical variables in the database S.

The (nominal) *dependence degree* of y on x is defined by

$$\begin{aligned}\omega^{y|x} &:= \sum_{i,j} p(y=i|x=j)p(x=j|y=i)p(y=i)\\ &= \sum_{i,j} p(x=j,y=i)p(y=i|x=j).\end{aligned}$$

Remarks.

N1 A measure of variance is referred to as *nominal* when any two distinct scenarios (or categories) are of equal distance.

N2 A measure of dependence degree of one variable on another is referred to as *nominal* if the involved variance is so. E.g.,

$$\text{Gini}(y) := \sum_i p(y=i)(1-p(y=i));$$

$$\text{Entropy}(y) := -\sum_i p(y=i)\log p(y=i).$$

N3 $\quad \omega^{y|x} \geq E(p(y)) = \sum_i p(y=i)^2.$

2. Feature Selection

The joint categorical distribution of several categorical variables can be regarded as a single nominal variable.

We are given a data set S with explanatory categorical variables

$$v_1, v_2, \ldots, v_n$$

and a response variable y with records

$$P_i := (x_{i1}, x_{i2}, \ldots, x_{in}, b_i), \qquad i = 1, \ldots, m.$$

Notation: $\mathsf{V}(1, 2, \ldots, n) := \{v_1, v_2, \ldots, v_n\}$.

Definition 2.1. A subset

$$V(i_1, i_2, \ldots, i_k) := \{v_{i_1}, v_{i_2}, \ldots, v_{i_k}\} \subseteq \mathsf{V}$$

is called an explanation base for y over S if

EB1. $\omega^{y|V(i_1,i_2,\ldots,i_k)} = \omega^{y|V(1,2,\ldots,n)}$;

EB2. for any $v \in V(i_1, \ldots, i_k)$,

$$\omega^{y|V(i_1,\ldots,i_k)\setminus\{v\}} < \omega^{y|V(1,2,\ldots,n)}.$$

Remark 2.2. There can be two or more explanation bases.

Proposition 2.3. If both $V(i_1, i_2, \ldots, i_k)$ and $V(j_1, j_2, \ldots, j_l)$ are bases for y on S, k can be equal or not equal to l, but both $|\mathrm{snr}(V(i_1, \ldots, i_k))|$ and $|\mathrm{snr}(V(j_1, \ldots, j_l))|$ are upbounded by

$$\max(\prod_{s=1}^{k} m_{v_{i_s}}, \prod_{t=1}^{l} m_{v_{j_t}}).$$

Proposition 2.4. For any given α with

$$E(p(y)) \leq \alpha \leq \omega^{y|V(1,2,\ldots,n)}, \qquad (*)$$

find a minimal subset of variables $V(i_1, \ldots, i_k)$ such that

$$\omega^{y|V(i_1,\ldots,i_k)} \geq \alpha. \qquad (**)$$

where the minimality means that if for any $v \in V(i_1, \ldots, i_k)$,

$$\omega^{y|V(i_1,\ldots,i_k)\setminus\{v\}} < \alpha.$$

Example 2.5. We have a categorical data set S consisting of 122 variables $v1, \ldots, v122$, and 24372 units. Here we take $v11$ to be a dependent variable y, of $y = v11$, and

$$E[P(y)] = \sum_{i=1}^{7} p(y = i)^2 = 0.232246,$$

where

y	1	2	3	4	5	6	7
p	.0005	.1317	.223	.3531	.1599	.1218	.01

With a forward-backward base variable selection procedure based on the measure of association $\omega^{y|x}$, we have obtained a base for y on S together with relative structure information as follows:

Cnt	var.	Gini	$\omega^{y\|v}$	Cumω	Scn	cumScn
1	v20	.7918	.3929	.3929	7	7
2	v22	.7770	.3806	.4483	7	37
3	v32	.7932	.3123	.4951	7	188
4	v13	.7764	.2718	.5706	6	773
5	v81	.7463	.2555	.6301	6	2605
6	v35	.7945	.2381	.7193	6	7088
7	v122	.7935	.2392	.8404	6	13920
8	v15	.7870	.2357	.9314	6	19549
9	v104	.7870	.2397	.9764	6	22512
10	v79	.7796	.2400	.9908	6	23547
11	v4	.7634	.2702	.9963	6	23956
12	v100	.7845	.2730	.9983	6	24140
13	v98	.6830	.2360	.9992	5	24213
14	v24	.6879	.2338	.9996	5	24264
15	v8	.7834	.2671	.9998	6	24296
16	v113	.7897	.2369	.9999	6	24310
17	v54	.5746	.2335	1	4	24326

With another forward-backward method, we get another base.

Cnt	Var.	Gini	$\omega^{y\|v}$	Cumω	Scn	cumScn
1	v20	.7918	0.3929	.3929	7	7
2	v22	.7770	0.3806	.4483	7	37
3	v32	.7932	0.3123	.4951	7	188
4	v9	.7752	0.2736	.5558	6	802
5	v17	.7879	0.3376	.5930	7	2351
6	v122	.7935	0.2392	.6825	6	6481
7	v104	.7870	0.2397	.8164	6	13158
8	v95	.7860	0.2649	.9119	7	18197
9	v35	.7945	0.2381	.9687	6	21917
10	v94	.7820	0.2636	.9865	7	23143
11	v19	.7690	0.2547	.9944	6	23795
12	v111	.7795	0.2873	.9969	7	24031
13	v1	.7874	0.2350	.9983	7	24188
14	v24	.6879	0.2338	.9991	5	24266
15	v34	.7887	0.3114	.9995	7	24306
16	v116	.5288	0.2339	.9998	4	24339
17	v15	.7870	0.2357	.9999	6	24353
18	v50	.7581	0.2346	1.000	6	24360