Rule Induction in Cascade Model Based on Sum of Squares Decomposition

Takashi Okada

Kwansei Gakuin University, Center for Information & Media Studies
Uegahara 1-1-155, Nishinomiya
662-8501 Japan
okada@kwansei.ac.jp

Abstract. A cascade model is a rule induction methodology using levelwise expansion of an itemset lattice, where the explanatory power of a rule set and its constituent rules are quantitatively expressed. The sum of squares for a categorical variable has been decomposed to within-group and between-group sum of squares, where the latter provides a good representation of the power concept in a cascade model. Using the model, we can readily derive discrimination and characteristic rules that explain as much of the sum of squares as possible. Plural rule sets are derived from the core to the outskirts of knowledge. The sum of squares criterion can be applied in any rule induction system. The cascade model was implemented as DISCAS. Its algorithms are shown and an applied example is provided for illustration purposes.

1 Introduction

The main subject of this paper is the explanatory power of a rule. Conventional rule induction systems give accuracy and coverage [1] or support and confidence [2] to a rule, but we cannot use a single measure for the explanatory power of a rule. Furthermore, we cannot know what portion of a problem has been solved by the rule.

The author proposed a cascade model [3] to solve this problem in discrimination rule induction. The model expanded the itemset lattice used in association rule mining [2, 4], where each item was expressed as a [attribute: value] pair of explanation attributes. The class distribution of instances was attached to each itemset in the lattice. A link in the lattice was employed as a rule:

IF *item-0* added on [*item-1*, *item-2*, …] THEN Class = *class-1*

where *item-0* is the item added along the link, and the other items in the LHS (left hand side) are those on the upper end of the link.

The latent discrimination power of the rule was defined as the product of the number of instances and the potential difference between the nodes,

$$\text{L-}power(U \rightarrow L) = N(L)\square(potential(U) - potential(L)), \qquad (1)$$

where *U* and *L* denote the upper and the lower nodes, respectively. The *gini*-index calculated from class distributions was employed as the potential of a node.

J.M. Zytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 468-475, 1999.

The total power of the problem was defined as a *root-node-power* in (2). Then, the explanatory power of the rule was quantitatively evaluated with respect to the *root-node-power*.

$$\textit{root-node-power} = N(\text{root}) \cdot \textit{potential}(\text{root}) \quad , \tag{2}$$

A view of cascades appeared when nodes and links are regarded as lakes and waterfalls, respectively. The problem of finding discrimination rules was transformed into the problem of selecting a set of powerful waterfalls explaining most of the *root-node-power*. The method was implemented as DISCAS software and successfully applied to an analysis of House voting records.

In this paper, we give more sound definitions to the power of a waterfall and to the potential of a lake based on sum of squares (*SS*). The next section gives a formulation of *SS* decomposition in the case of a categorical variable. Section 3 discusses the rule expression. Algorithms for the selection of rules are described in Sect. 4, and the results of an application to a simple illustrative dataset are shown in Sect. 5.

2 Decomposition of Sum of Squares

It is well known that the total sum of squares (*TSS*) of a continuous variable can be decomposed to the sum of within-group sums of squares (WSS^g) and between-group sums of squares (BSS^g) by the following equation [5]. If we derive the same equation for a categorical variable, BSS^g will be suitable as the power definition of a link to group *g*.

$$TSS = \sum_g \left(WSS^g + BSS^g\right) \quad , \tag{3}$$

Following the description in [6], Gini showed that the *SS* definition for *x* could be transformed to (4), and that the distance definition of a categorical variable by (5) led to the expression of *SS* by (6),

$$SS = \frac{1}{2n}\sum_a \sum_b (x_a - x_b)^2 \quad , \tag{4}$$

$$x_a - x_b \begin{cases} =1 & \text{if } x_a \neq x_b , \\ =0 & \text{if } x_a = x_b , \end{cases} \tag{5}$$

$$SS = \frac{n}{2}\left(1 - \sum_\alpha p(\alpha)^2\right) \quad , \tag{6}$$

where *n* shows the number of instances, *a, b* each designate an instance, and $p(\alpha)$ is the probability of class α. The formula in parentheses in (6) is the *gini*-index. This definition of *SS* leads to the expressions of *TSS* and WSS^g in (7) and (8),

$$TSS = \frac{n}{2}\left(1 - \sum_\alpha p(\alpha)^2\right) \quad , \tag{7}$$

$$WSS^g = \frac{n^g}{2}\left(1 - \sum_{\alpha} p^g(\alpha)^2\right) \quad . \tag{8}$$

If we give the definition (9) to BSS^g, the decomposition of *TSS* in (3) is shown to be valid [7]. Thus, we can employ BSS^g as the latent power of a link in cascades. Furthermore, the sample variance of a group, the quotient of WSS^g by n^g, can be considered as the potential of a node.

$$BSS^g = \frac{n^g}{2}\sum_{\alpha}\left(p^g(\alpha) - p(\alpha)\right)^2 \quad , \tag{9}$$

An example of *SS* decomposition is shown in Fig.1, where *TSS*: 160 is decomposed to *BSS* and *WSS* for the groups 1 and 2. The class distribution of the lower right group is opposite to that of the upper node, and hence the link between these nodes obtains a large BSS^g value, as expected.

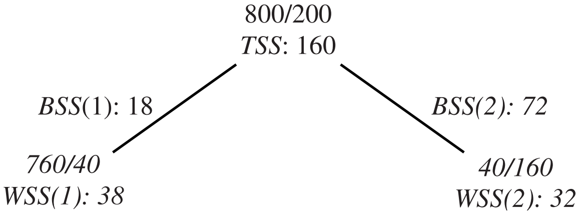

Fig. 1. Example of *SS* decomposition. The top line at each node shows numbers of instances for the two classes

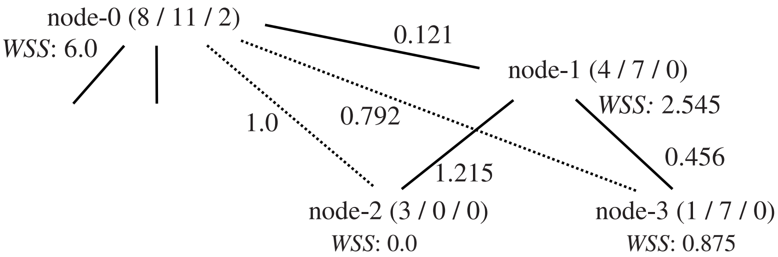

Fig. 2. Example of *SS* conservation. Numbers of instances for 3 classes are shown at each node. Values attached to links denote *BSS* values

Another interesting example of *SS* decomposition is shown in Fig. 2. A part of *SS* at node-0 is decomposed to node-1, of which *WSS* is again divided to nodes-2, 3. Another possible decomposition is shown by the dotted lines, where the same *SS* is directly decomposed into nodes-2, 3. In both cases, summation of the relevant *BSS*s gives the same value: 1.792. However, addition of the *BSS*s between nodes -1, 2 and between nodes -0, 3 gives 2.007. This result indicates that conservation of *SS* is guaranteed only when *BSS*s are taken from the links in a tree.

SS decomposition can be applied recursively. Then, the *SS* at the root node of a decision tree is decomposed into the *BSS*s of all the links and into the *WSS*s of all the terminal nodes. Hence, the *SS* at the root node may be regarded as the total power of the problem to be solved.

3 Rule Expression

Let us suppose a rule link between nodes U and L as shown in Fig. 3. The itemset on U consists of items [A: a1], [B: b1], and the item [C: c1] is added along the link. The attributes X, Y, which do not appear in these itemsets, are called veiled attributes at the node. The items derived from veiled attributes are called veiled items. The distributions of these veiled items are attached at the right of each node.

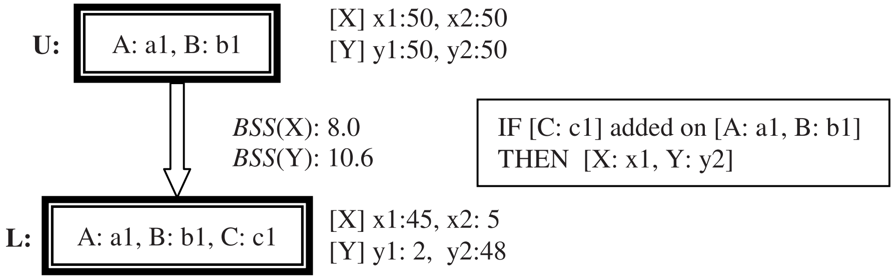

Fig. 3. Link, distributions and *BSS*s of veiled items, and the resulting rule expression

The previous section introduced the *SS* decomposition for the class attribute, which is a veiled attribute. However, there are no obstacles to applying the scheme to other veiled attributes. In fact, we can compute the *BSS* of *X* and *Y*, as shown in Fig. 3. In this example, the percentage of veiled items [X: x1] and [Y: y2] shows a sharp increase along this link, giving high *BSS* values to these attributes. We can say that the interactions between the added item [C: c1] and the veiled items [X: x1], [Y: y2] are large in the extension of node U. This recognition of strong interaction leads to the rule expression in the right textbox in Fig. 3.

The rule expression is the same as an association rule, if we merge the items in the LHS of the rule. However, an association rule miner generates a rule by a comparison of two itemsets: [A: a1, B: b1, C: c1] and [A: a1, B: b1, C: c1, X: x1, Y: y2], whereas the current method compares two itemsets, [A: a1, B: b1] and [A: a1, B: b1, C: c1] to describe the LHS, and the items in the RHS are derived from the sample distribution of veiled items along the rule link.

Selection of items in the RHS consists of two steps. First, we choose veiled attributes with *BSS* values that exceed a given parameter. Then, we select a value, α, of the attribute that corresponds to the maximum term in the summation of (9). If there are plural values with the same contribution, the value with the largest $p^g(\alpha)$ is selected. All [attribute: value] pairs thus derived are written as items in the RHS.

The selection of rules based on *BSS* values allows recognition of negative interactions among items, that is, the percentage of instances may decrease along a link, which it is also important to know.

4 Extraction of Rules

We have developed DISCAS (version 1.3) software that can detect interactions using the *BSS* of a class variable and of explanatory variables. The resulting rules are called discrimination rules and characteristic rules, respectively. An efficient lattice generation algorithm appears separately [8]. Here we show two key concepts in the extraction of rules.

Selection of Candidate Rule Links by Voting. We associate an instance to a link, if the item added along the link is contained in the instance. An instance selects the l links with the highest *BSS* values among the associated links. The selected links receive a vote from this instance. Links with at least m votes are chosen as candidate rule links. Default values of l and m were set to 3 and 1.

As for a *BSS* value, we use that of the class attribute when we generate discrimination rules. The sum of the *BSS* values for all veiled explanation attributes was used to generate characteristic rules.

Plural Rule Sets and Rules Selection. Rules are expressed as a group of rule sets. The rules in a rule set are selected so that their supporting instances cover all instances.

We have employed a greedy algorithm to extract rules from candidate links. That is, the link with the highest *BSS* value is selected as the first rule in the first rule set. Its supporting instances are deleted from the support of the remaining candidate links. This process is successively applied to the candidates, until there are no candidates with supporting instances. The rules selected constitute the first rule set.

After a rule set is created, we recover the initial supporting instances for the remaining candidate links, and creation of the next rule set proceeds as for the first. The rule induction process is complete when all the candidates are expressed as rules.

5 Application to Cars Dataset

DISCAS has been applied to a simple cars dataset to illustrate the capability of the cascade model. This data set consists of only 21 instances and 10 categorical attributes [9]. The problem is to predict the mileage of cars from the other 9 attributes. The entire lattice was created except the itemsets containing mileage attributes. Default values are used for the selection of rule links. The first rule sets for discrimination and characteristic rules are shown in Tables 1 and 2, respectively.

Discrimination Rules. The eight rules in Table 1 are ordered according to their *BSS* values. For example, we should read the fifth rule as follows, IF [cyl: 6] added on [fuelsys: EFI, trans: manual] THEN [mileage: high].

Vowels are omitted from attribute names in the table. The number of supporting instances and the confidence levels are 11 (54.5%) and 4 (0%) on the upper and the lower nodes, respectively. The three *BSS* values have the following meanings. The L-*BSS* (latent *BSS*) column shows the *BSS* values defined in Sect. 2. Of the supporting instances for the fifth rule, two have already been used for the third and the fourth rules, and hence the number of effective instances remaining is 2. The E-*BSS* (effective *BSS*) column shows the effective part of the L-*BSS*. The R-*BSS* (*BSS* from root node) column denotes the *BSS* of a link from the root node to the virtual node consisting of these effective instances. We can sum R-*BSS* in the rule set, and compare the value with *SS* at the root node. The RHS of this rule should be read as [mileage: not high], because its confidence level decreases to 0.0%.

Table 1. Discrimination rules from cars example (see text for explanation of terms used)

No	Item added Items on upper node	RHS	confidence (%)	support	*BSS*		
					L	E	R
1	Sz: subcompact Null	mlg: high	38.1 : 100	21 : 6	2.00	2.00	2.00
2	Wght: heavy Null	mlg: low	9.5 : 100	21 : 2	1.24	1.24	1.24
3	cyl: 6 flsys: EFI; cmp: high	mlg: medium	28.6 : 100	7 : 2	1.02	1.02	0.38
4	sz: compact cyl: 4; trb: no	mlg: medium	45.5 : 83.3	11 : 6	0.86	0.86	0.45
5	cyl: 6 flsys: EFI; trns: manual	mlg: high	54.5 : 0.0	11 : 4	0.94	0.47	0.38
6	flsys: EFI sz: compact; cyl: 4; cmp: high	mlg: high	40.0 : 100	5 : 2	0.72	0.36	0.33
7	trns: auto flsys: EFI	mlg: medium	50.0 : 100	14 : 3	0.66	0.22	0.19
8	cmp: medium cyl: 4; flsys: EFI	mlg: medium	25.0 : 66.7	8 : 3	0.52	0.17	0.19

Table 2. Characteristic rules from cars example (see text for explanation of terms used)

No	item added items on upper node	RHS	confidence (%)	support	*BSS*		
					L	R	sum
1	dsplc: small null	cyl: 4 cmp: high	66.7 : 100 52.4 : 88.9	21 : 9	1.00 1.20	1.00 1.20	7.27
2	pwr: high trb: no	cyl: 6 dsplc: medium	35.3 : 80.0 52.9 : 100	17 : 5	1.00 1.11	1.09 0.92	5.14
3	Dsplc: medium Null	-	-	21 : 12	-	-	2.57

The *SS* of the mileage at the root node is 6.0, while the sum of R-*BSS* and E-*BSS* of the rules in the rule set are 5.17 and 6.34, respectively. The difference (0.83) between the sum of R-*BSS* and *SS* at the root node is equal to the *WSS* of the lower node of rule-4.

The top five rules in this rule set cover 18 out of all 21 instances, and their R-*BSS* explains 74% of the *SS* at the root node. We do not need to worry about numerous rules, because we can stop rule inspection when the sum of the *BSS* values reaches close to the *SS* at the root node. If our interest in a discrimination mechanism is from a different viewpoint, we can browse rules with high latent *BSS* values in the succeeding rule sets.

Characteristic Rules. The first rule set among the explanatory attributes consists of the three rules in Table 2. The RHS column shows the veiled items where L-*BSS* exceeds 1.0. The column at the right end shows the sum of the R-*BSS* values of all items except the class attribute: mileage. In rule-3, there are no items with high L-*BSS*.

The sum of all R-*BSS* in this rule set is 15.0. We can say that 36% of the *SS* at the root node (41.81) is covered by these 3 links. This suggests a moderate correlation among the 9 explanatory attributes. On the other hand, the sum of R-*BSS* values for the class variable: mileage explains only 20% of its *SS* at the root node in these 3 rules. If our object is the segmentation of cars, these 3 links will lead to good starting clusters, but they are not necessarily good for mileage prediction.

6 Related Works

In statistics, considerable effort has been directed to detecting interactions among variables [10] and to generating decision trees [1]. Many clustering methods use the sum of squares criterion for continuous variables [5]. However, there has been no attempt to use the sum of squares definition of categorical variables as a criterion in tree formation or in clustering.

Levelwise construction of lattices is now an accepted method [2, 4], and there have been attempts to extract classification rules using association rule miners [11, 12, 13]. Many measures to select rules have already been discussed [14]. Among them, our method has common conceptual points with maximal guess [15] and exceptional knowledge discovery [16]. Detection of interactions using the χ^2 statistic is especially relevant to the current work [17]. However, no research to date has recognized the importance of *SS*.

7 Concluding Remarks

Decomposition of *SS* has been proposed to give a criterion of rule power in the cascade model. The DISCAS system has been applied to the induction of discrimination and characteristic rules, giving satisfactory results. The method is expected to be a powerful tool for data analysis. Further, we have shown that the RHS of a rule can be found from the distribution of veiled items at the LHS nodes. This finding has led to an efficient pruning methodology [8].

Algorithms employed are subject to change, but the *SS* criterion can easily be adapted to other systems and will provide a useful measure. *SS* is one of the core concepts in statistics, so bridging between statistics and rule induction is expected.

Acknowledgements

The author would like to thank Prof. Sugihara, Prof. Oyama and Dr. Jimichi of Kwansei Gakuin University for their encouragement and their valuable discussions about statistics.

References

1. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth, Belmont CA (1984)
2. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. Proc. ACM SIGMOD (1993) 207-216
3. Okada, T.: Finding Discrimination Rules Based on Cascade Model. submitted to J. Jpn. Soc. Artificial Intelligence (1999)
4. Agarawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. Proc. VLDB (1994) 487-499
5. Takeuchi, K. et al. (eds.): Encyclopedia of Statistics (in Japanese). Toyo Keizai Shinpou (1990) 388-390
6. Light, R.J., Margolin, B.H.: An Analysis of Variance for Categorical Data. J. Amer. Stat. Assoc. 66 (1971) 534-544
7. Okada, T.: Sum of Squares Decomposition for Categorical Data. to appear in Kwansei Gakuin Studies in Computer Science 14 (1999). http://www.media.kwansei.ac.jp/home/kiyou/kiyou99/kiyou99-e.html
8. Okada, T.: Efficient Detection of Local Interactions in Cascade Model. submitted to Discovery Science 99
9. Ziarko, W.: The Discovery, Analysis, and Representation of Data Dependencies in Databases. In: Piatetsky-Shapiro, G., Frawley, W. (eds.): Knowledge Discovery in Databases. AAAI Press (1991) 195-209
10. Hawkins, D.M., Kass, G.V.: Automatic Interaction Detection. In: Hawkins D.M. (ed.): Topics in Applied Multivariate Analysis. Cambridge Univ. Press (1982) 269-302
11. Ali, K., Manganaris, S., Srikant, R.: Partial Classification using Association Rules. Proc. KDD-97 (1997) 115-118
12. Bayardo, R.J.: Brute-Force Mining of High-Confidence Classification Rules. Proc. KDD-97 (1997) 123-126
13. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. Proc. KDD-98 (1998) 80-86
14. Silberschatz, A. and Tuzhilin, A.: What Makes Patterns Interesting in Knowledge Discovery Systems. IEEE Trans. on Know. and Data Eng. 8(6) (1996) 970-974
15. Washio, T., Matsuura, H., Motoda, H.: Mining Association Rules for Estimation and Prediction. In: Wu, X., Kotagiri, R., Korb, K.B. (eds.) Research and Development in Knowledge Discovery and Data Mining. Lecture Notes in A.I. Vol. 1394. Springer-Verlag (1998) 417-419
16. Suzuki, E., Shimura, M.: Exceptional Knowledge Discovery in Databases based on Information Theory. Proc. KDD-96 (1996) 275-278
17. Silverstein, C., Brin, S., Motwani, R.: Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. Data Mining and Knowledge Discovery, 2 (1998) 39-68