Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 76(2) 2022 85–92

Application of STUCCO Algorithm for Finding Contrast Sets for Agricultural Datasets

Sonica Priyadarshini[1], Alka Arora[1], Rajni Jain[2], Sudeep Marwaha[1], Anshu Bharadwaj[1], A.R. Rao[3] and Soumen Pal[1]

[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi
[2]ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi
[3]Indian Council of Agricultural Research, New Delhi

Received 25 May 2022; Revised 17 September 2022; Accepted 23 September 2022

SUMMARY

The interplay between computer science and agriculture has led to the collection of huge amounts of information in agricultural datasets. The process of turning low level data into high level knowledge is popularly known as data mining. The field of agriculture has many applications and one important application is in terms of deriving useful patterns like characteristics of disease and varieties. Understanding the distinctions between numerous contrasting groups is a crucial issue in data analysis in order to discover new patterns. These contrasting groups can represent various item classes, such as disease or varieties for different crop groups. Contrast sets are the combinations of attributes and their values that differ meaningfully in their distribution across groups. STUCCO algorithm is a search method for mining contrast sets leading to pattern discovery. The algorithm's applicability for pattern detection has been demonstrated using agricultural datasets in this paper. Approach resulted in significant pattern discovery for description of soybean disease and IRIS varieties characteristics.

Keywords: Contrast set, Pattern discovery, Statistical significance, STUCCO algorithm.

1. INTRODUCTION

Data Mining is a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data (Han and Kamber, 2006). The underlying assumption of data mining is to find out the hidden patterns in the data, which can be revealed by grouping the objects into classes. Producing a pattern is of interest in the situation where there is a need to study the relationship describing the data. Pattern discovery can be applied in various areas for understanding the patterns viz- disease diagnostic system (to study the diseases characteristics), Web Mining (to find pattern in the set of web users), tourism industry (to find what features of places and tourist attract each other), banks (to identify defaulters) and agriculture (to characterize animal & plant diseases and variety).

Bay and Pazzani (2001) proposed Contrast Set Mining as a technique to identify significant differences among the groups. Bay and Pazzani introduced the STUCCO algorithm for finding contrast sets. Contrast sets are conjunctions of attributes and values pairs that differ meaningfully in their distributions across groups. The contrast set is a stepwise computation of support and significance test. In order to improve classification accuracy and minimize required time, contrast set mining is used in feature selection and pattern discovery.

There are some published studies related to pattern recognition and contrast set mining concepts. Kralj *et al.* (2007) worked on an approach to the subgroup discovery task. He was able to successfully apply the method to the study of records of patients with brain stroke. Novak *et al.* (2009) surveyed Contrast Set

Corresponding author: Sonica Priyadarshini
E-mail address: priyadarshini.bhu.10@gmail.com

Mining (CSM), Emerging Pattern Mining (EPM), and Subgroup Discovery (SD) in the context of supervised descriptive rule discovery. A critical survey was conducted for existing supervised descriptive rule discovery visualization techniques. Langohr *et al.* (2013) mentioned that subgroup discovery methods find interesting subsets of objects of a given class. Contrast set mining, according to Magalhes and Azevedo (2009), is based on identifying significant patterns by contrasting two or more groups. They also defined a set of temporal patterns to represent the significant changes in contrasts identified across the time period under consideration. Boettcher (2011) explained and compared contrast set mining and change mining. He mentioned that the contrast set describes what changes are there, in terms of differences while change mining is a data-mining paradigm for the study of time-associated data. Kaneiwa *et al.* (2011) has explained about sequential pattern mining. They have used rough set theory for finding the decision rules and developing a sequential information system. Qian *et al.* (2020) applied the concept of contrast set mining on facebook data for pattern mining.

The concept of contrast set has never been applied on any agricultural datasets for pattern discovery. This paper demonstrates the applicability of contrast set mining for pattern detection using agricultural datasets.

2. CONTRAST SET

2.1 Definition

The data is a set of k-dimensional vectors where each component can take on a finite number of discrete values. The vectors are organized into "n" mutually exclusive groups $G_1, G_2, \ldots, G_n$, with $G_i \cap G_j = \varnothing$ $\forall i \neq j$. Let $A_1, A_2, \ldots, A_k$ be a set of k variables called attributes. Each A_i can take on values from the set $\{V_{i1}, V_{i2}, \ldots, V_{im}\}$. Then a contrast set is a conjunction of attribute-value pairs defined on groups $G_1, G_2, \ldots, G_n$ with no A_i occurring more than once (Bay and Pazzani, 1999). In soybean crop, External decay = Firm & dry Λ Temperature < Normal identifies Rhizoctonia Root Rot disease and Sclerotia = Present Λ Canker Lesion = Tan identifies Charcoal rot disease; both are examples of contrast set.

The support of a contrast set with respect to a group G is the percentage of examples in G where the contrast set is true. The main goal is to find all contrast sets whose support differs meaningfully across groups (Bay and Pazzani, 1999). Contrast sets are usually denoted as cset or c and support is denoted as P(cset | G) or support(cset, G).

$$\text{Max } i\ j|\text{support}(\text{cset}, G_i) - \text{support}(\text{cset}, G_j)| \geq \delta \qquad (1)$$

$$\exists i\ j\ P(\text{cset} = \text{True} \mid G_i) \neq P(\text{cset} = \text{True} \mid G_j) \qquad (2)$$

And δ is a threshold called the minimum support difference which is user defined. Large contrast sets are those that meet Eq. (1), whereas significant contrast sets are those that meet Eq. (2) statistically. When both prerequisites are satisfied, it is referred to as a deviation. The first criterion measures the effect size and ensures that everything reported as a result is a big enough effect to be important. The statistical significance requirement guarantees that the contrast set accurately depicts the differences between groups.

2.2 STUCCO Algorithm

Bay & Pazzani (1999) introduced STUCCO (Search and Testing for Understandable Consistent Contrasts), with the benefits of a pruning mechanism. It uses a breadth-first search approach, which incorporates several techniques from work on efficiently mining large datasets.

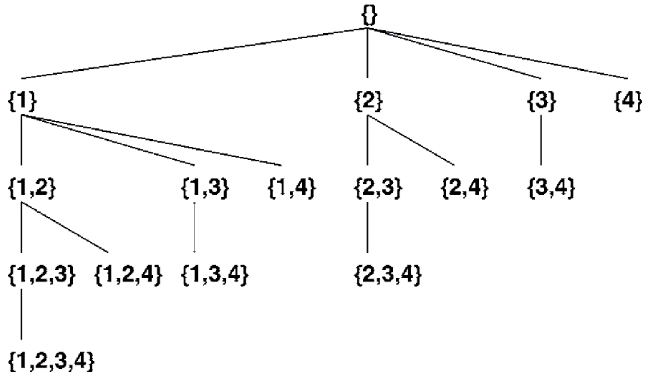


Fig. 1. Example search tree for four attribute-values pairs with ordering {1,2,3,4}

The search for contrast sets was organized using set-enumeration trees (Rymon, 1992; Bayardo,1998) to ensure that every node is visited only once or not at all if nodes can be pruned. Breadth-first search is used because it proceeds in a level-wise manner (Fig. 1). It means one can go through all attributes separately in first level then all possible conjunctions of two attributes in second level and so on. The level-wise nature allows to present results in an anytime fashion. At each level of the search, the database is scanned and the support is counted of all nodes for each group. The support counts were examined to determine which nodes meet our criteria and which nodes should be pruned and then moved to the next level (Bay and Pazzani, 1999).

2.2.1 Support

The Support of a contrast set is calculated with respect to a group Gi as:

$$S\,(\%) = n \text{ X } 100/N \tag{3}$$

Where n = number of observations for which the contrast set is true

N = total number of observations

Support difference is being calculated across the class (disease, variety etc.). The minimum deviation or minimum support difference (δ) is a user-defined criterion. A large set is one in which the support difference is higher than or equal to the minimum support difference. We set the minimum deviation value to 100% in order to achieve better and more accurate findings. As a result, only those attribute value pairs having strong support were chosen. After finding the 'large' attribute value pair, a test of significance was done.

2.2.2 Significance test

Chi-square (χ2) statistic is used to determine the significance of contrast sets, which are a large set. For testing the equality of contrast set support across all groups, a null hypothesis was considered. Level of significance (α) was taken 5%. By taking the row variable as the truth of the contrast set, and the column variable as the group membership, a 2 × *G* contingency table was formed. The chi-square test is the standard test for variable independence in contingency tables. It works by computing the statistic χ2:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4}$$

Where *Eij* = expected frequency count in cell *ij* given independence of the row and column variables

Oij = observed frequency count for the cell in row *i* and column *j*

c = Total number of classes/columns present in data

Eij is calculated as follows:

$$E_{ij} = \frac{\boxed{\sum_{i=1}^{2} O_{ij}}\boxed{\sum_{j=1}^{c} Oij}}{N} \tag{5}$$

Where *N* = total number of observations.

With the help of a chi-square table, the comparison of results was done with the distribution of χ2 when the null hypothesis is true (Annexure I).

3. RESULTS AND DISCUSSION

In this section, the application of contrast set mining on soybean and iris datasets (UCI Repository) have been illustrated. Soybean disease set contains 47 observations and set of attributes consist of 35 multi-valued variables that characterizes 4 diseases: diaporthe-stem-canker (D1), charcoal-rot (D2), rhizoctonia-root-rot (D3) and phytophthora-rot (D4). All the variables are nominal in nature. Variables are broadly categorized into environmental descriptors, condition of leaves, condition of stem, condition of fruit pods and condition of root. It is observed that the dataset is having unique value for some of the variables hence those variables are irrelevant and removed from the dataset during data cleaning. Reduced dataset has 21 variables that characterize soybean diseases (Annexure II).

As mentioned earlier, this algorithm is based on breadth-first search approach. Therefore, in the first level all 21 attributes were taken individually to see if any of them fall into the contrast set. The support was

Table 1. Calculation of support

Attribute Value	D1	D2	D3	D4	Max-Min Support Difference	Remarks
Precipitation < Normal	0	10 X 100/10 = 100	0	0	100 - 0 = 100	Large
Precipitation = Normal	0	0	0	4 X 100/17 = 23.53	23.53 - 0 = 23.53	-
Precipitation > Normal	10 X 100/10 = 100	0	10 X 100/10 = 100	13 X 100/17 = 76.5	100 - 0 = 100	Large
Temperature < Normal	0	0	10 X 100/10 = 100	7 X 100/17 = 41.18	100 - 0 = 100	Large
Temperature = Normal	10 X 100/10 = 100	4 X 100/10 = 40	0	10 X 100/17 = 58.8	100 - 0 = 100	Large
Temperature < Normal	0	6 X 100/10 = 60	0	0	60 - 0 = 60	-

calculated for all attribute value pairs. Then the Support difference was calculated. For getting better and more accurate results, the minimum deviation value was taken as 100%. Therefore, only those attribute value pairs were selected having 100 % support difference (Table 1).

From the above mentioned method the 'large' attribute value pairs were found, and then a significance test was executed. A *2 X 4* contingency table was prepared (Table 2(a)) followed by the expected value table (Table 2(b)).

Table 2 (a). Contingency table for "Precipitation < Normal"

	D1	**D2**	**D3**	**D4**	**Total**
c	0	10	0	0	10
¬ c	10	0	10	17	37
Total	10	10	10	17	47

Table 2 (b). Expected values for "Precipitation < Normal"

	D1	**D2**	**D3**	**D4**	**Total**
c	0	10	0	0	10
E (c)	2.13	2.13	2.13	3.62	
¬ c	10	0	10	17	37
E (¬ c)	7.87	7.87	7.87	13.38	
Total	10	10	10	17	47

$$\chi^2 =$$

$$(0 - 2.13)^2 / 2.13 + (10 - 2.13)^2 / 2.13 + (0 - 2.13)^2 / 2.13 + (0 - 3.62)^2 / 3.62 + (10 - 7.87)^2 / 7.87 + (0 - 7.87)^2 / 7.87 + (10 - 7.87)^2 / 7.87 + (17 - 13.38)^2 / 13.38 = 47.0$$

The degree of freedom of the *R X C* contingency table is *(R-1)(C- 1)* so for the *2 X 4* contingency table degree of freedom (d.f.) is (2-1)(4-1) = 3. For three degrees of freedom, a χ^2 value larger than 7.82 is taken as significant according to the chi square table. It indicates that Precipitation < Normal is a significant attribute value pair. Therefore, it is a contrast set. Similarly the algorithm works on the whole dataset.

After applying the algorithm, 25 contrast sets were found which can differentiate 4 diseases (Table 3). Among these 25 contrast sets 8 sets were most significant and can differentiate diseases uniquely with 100% accuracy (Table 4 and Fig. 2, 3).

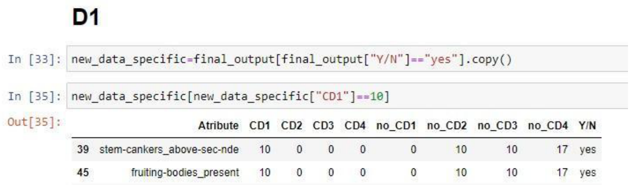
In the second level, all possible combinations of pairs from all 21 attributes were taken under consideration for finding a contrast set. There were 386 contrast sets which can differentiate all 4 diseases up to some extent. Overall, 37 contrast sets can uniquely differentiate diaporthe-stem-canker from other diseases, 24 contrast sets among them can uniquely differentiate diaporthe-stem-canker from others with 100% accuracy. There are 77 contrast sets that can uniquely separate charcoal-rot from other diseases, with 53 of them being able to do so with the accuracy of 100%. The STUCCO algorithm was implemented in python language using jupyter notebook. Python is a multi-purpose, high-level programming language which is being widely used for data analysis. It allows programming in object-oriented and procedural paradigms. Mainly two libraries were used for applying this algorithm on python - NumPy and Pandas. Fig. 2 and Fig. 3 show the screenshots of the results for Diaporthe Stem Canker (Fig. 2) and Charcoal Rot (Fig. 3) from the developed software.

Table 3. Contrast set with single attribute for Soybean disease

Plant Stand = Normal	Plant Stand < Normal
Precipitation < Normal	Precipitation > Normal
Temperature < Normal	Temperature = Normal
Area Damaged = Lower Areas	Stem Canker = Absent
Stem Canker = Below Soil	Stem Canker = Above Second Node
Canker Lesion = Brown	Canker Lesion = Dark Brown-Black
Canker Lesion = Tan	Fruiting Bodies = Absent
Fruiting Bodies = Present	External Decay = Absent
External Decay = Firm & Dry	Initial Discoloration = None
Initial Discoloration = Black	Sclerotia = Absent
Sclerotia = Present	Fruit Pods = Normal
Fruit Pods = dna	Roots = Normal
Roots = Rotted	

Table 4. Most significant Contrast set with single attribute for Soybean disease

Most Significant Contrast Set	**Uniquely Differentiating Disease**
Stem Canker = Above Second Node	Diaporthe stem canker (D1)
Fruiting Bodies = Present	Diaporthe stem canker (D1)
Precipitation < Normal	Charcoal rot (D2)
Stem Canker = Absent	Charcoal rot (D2)
Canker Lesion = Tan	Charcoal rot (D2)
Initial Discoloration = Black	Charcoal rot (D2)
Sclerotia = Present	Charcoal rot (D2)
Canker Lesion = Dark Brown-Black	Phytophthora rot (D4)



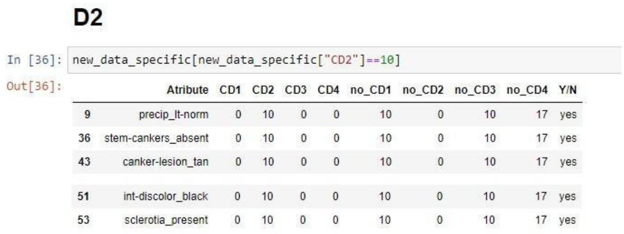
D1

```
In [33]: new_data_specific=final_output[final_output["Y/N"]=="yes"].copy()

In [35]: new_data_specific[new_data_specific["CD1"]==10]

Out[35]:
                       Atribute  CD1  CD2  CD3  CD4  no_CD1  no_CD2  no_CD3  no_CD4  Y/N
39  stem-cankers_above-sec-nde   10    0    0    0       0      10      10      17  yes
45       fruiting-bodies_present   10    0    0    0       0      10      10      17  yes
```

Fig. 2. Screenshot of result for Diaporthe stem canker



D2

```
In [36]: new_data_specific[new_data_specific["CD2"]==10]

Out[36]:
               Atribute  CD1  CD2  CD3  CD4  no_CD1  no_CD2  no_CD3  no_CD4  Y/N
 9        precip_lt-norm    0   10    0    0      10       0      10      17  yes
36   stem-cankers_absent    0   10    0    0      10       0      10      17  yes
43      canker-lesion_tan    0   10    0    0      10       0      10      17  yes

51      int-discolor_black    0   10    0    0      10       0      10      17  yes
53       sclerotia_present    0   10    0    0      10       0      10      17  yes
```

Fig. 3. Screenshot of result for Charcoal rot

Overall, 23 contrast sets can differentiate rhizoctonia-root-rot from others uniquely, 5 contrast sets among them can uniquely differentiate rhizoctonia-root-rot from other diseases with accuracy of 100% and 5 contrast sets with accuracy of 90%. There are 23 contrast sets that can uniquely separate phytophthora-rot from other diseases, with 10 of them able to do so with 100% accuracy and one with 94% accuracy.

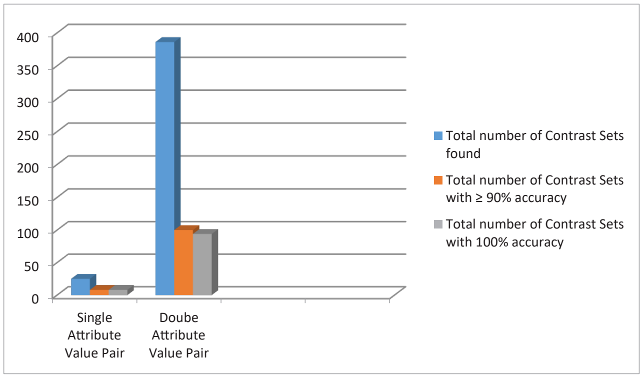


Fig. 4. Graphical representation of extracted Contrast Sets for single and double attributes for Soybean disease dataset

Total 25 contrast sets were found with a single attribute value pair which can differentiate all 4 diseases. Among these 25 contrast sets, there were 8 contrast sets which can differentiate diseases uniquely with 100% accuracy (Fig. 4). There are 2 contrast sets that can separate diaporthe-stem-canker uniquely from other diseases. Overall, 5 contrast sets can separate charcoal-rot uniquely from other diseases. And 1 contrast set can separate phytophthora-rot uniquely from other diseases.

Total 386 contrast sets were found at the second level with double attribute value pairs which can differentiate all 4 diseases. Among them 94 contrast sets were most significant and they were able to separate diseases uniquely with 100% accuracy (Fig. 4). There were 24 contrast sets that can separate diaporthe-stem-canker diseases uniquely from others. Overall, 55 contrast sets were found that can separate charcoal-rot disease uniquely from others. Among them 5 contrast sets can separate rhizoctonia-root-rot disease uniquely from others. Among them 10 contrast sets can separate phytophthora-rot disease uniquely from others.

The above-mentioned findings were compared to those of Arora *et al.* (2009a) and Jain *et al.* (2013). A Reduct Driven Cluster Description (RCD) approach was applied on the Soybean dataset for the selection of significant variables from individual clusters by Arora *et al.* (2009a)**.** Jain *et al.* (2013) applied Multiple Pattern Formulation approach for pattern discovery in Soybean dataset. The comparison was done for obtained contrast sets for single and double attributes from the STUCCO algorithm. There was similarity in almost all results or patterns found with Arora *et al.* (2009a) and Jain *et al.* (2013). These findings were also cross checked with the symptoms explained by Hartman *et al.* (1999) and Gupta *et al.* (2005). The majority of the symptoms mentioned by them were identified correctly. According to Hartman *et al.* (1999) and Gupta *et al.* (2005) fields with a notable incidence of stem canker may be detected at any time from flowering through pod fill in case of diaporthe stem canker. Seed develops black discolouration in case of charcoal rot. Dry conditions, relatively low soil moisture and nutrients and temperature ranging from 25^{o}C to 35^{o}C are favourable for the disease. Production of abundant minute black sclerotia beneath the outer cortical tissues and in the pith region, which turn to silvery white to light black, is a diagnostic symptom of the charcoal rot disease. The stem rot phase in phytophthora rot disease is easily recognizable by the presence of a distinct chocolate-brown lesion moving up the stem from the soil line. All above mentioned symptoms were found as significant contrast sets and mentioned in Table 4.

Similarly, the contrast set mining was applied for the iris dataset. Before applying contrast set mining each attribute value of iris data was discretized into 3 categories (Annexure III). For getting better and more accurate results, the minimum deviation value was set to 50%. At first level 11 contrast sets were found among which 2 contrast sets were able to differentiate the varieties with 100% accuracy. There were 8 contrast sets that were able to differentiate the varieties with more than 80% accuracy (Table 5).

Table 5. Contrast set with single attribute for Iris data

Attribute	Attribute-value (in cms)	Variety	Accuracy
Sepal Length	Greater than 6.7	Iris-virginica	85%
Sepal Width	Greater than 3.6	Iris-setosa	86.67%
Petal Length	Less than 2.97	Iris-setosa	100%
Petal Length	2.97 – 4.93	Iris-versicolor	88.89%
Petal Length	Greater than 4.93	Iris-virginica	95.65%
Petal Width	Less than 0.9	Iris-setosa	100%
Petal Width	0.9 – 1.7	Iris-versicolor	90.7%
Petal Width	Greater than 1.7	Iris-virginica	97.83%

At second level 42 contrast sets were found among which 21 contrast sets were able to differentiate the varieties with 100% accuracy. There were 38 contrast sets that are able to differentiate the varieties with more than 80% accuracy (Fig. 5). The above results were compared with the study conducted by Arora *et al.* (2009b) and were found similar.

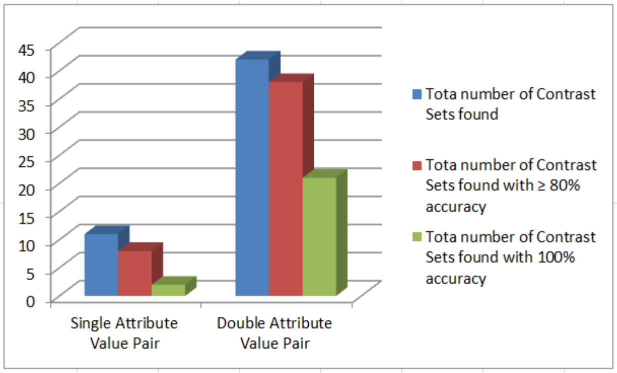


Fig. 5. Graphical representation of extracted Contrast Sets for single and double attributes for Iris dataset

4. CONCLUSION AND FUTURE SCOPE

Contrast set mining aids in the identification of a substantial list of attribute value pairs that differ significantly from others. As a result, it aids pattern discovery and feature selection. In this work, we discussed how a contrast set assisted in identifying the disease-causing characteristics. This method yielded 25 contrast sets for a single characteristic, with 8 being the most significant for the soybean disease dataset. For double attributes, 386 contrast sets were retrieved, with 98 being the most significant for the identical data. With the same approach 11 contrast sets for a single attribute were extracted for the iris dataset, among which 2 were most significant. At second level 42 contrast sets were found among which 21 was most significant for the same data. Similar concepts can be used for variety characterization, learning rules from data for expert systems etc. So, there is scope of work in continuation to the mentioned approach in future. In future, different approaches can be studied to apply pattern discovery for continuous dataset as well as to enhance the efficiency of algorithms.

REFERENCES

Arora, A., Upadhyaya, S. and Jain, R. (2009a). Post Processing of Clusters for Pattern Discovery: Rough Set Approach, *Journal of the Indian Society of Agricultural Statistics*, **63(2)**, 181-188.

Arora, A., Upadhyaya, S. and Jain, R. (2009b). Integrated Approach of Reduct and Clustering for Mining Patterns from Clusters, *Information Technology Journal*, **8(2),** 173-180.

Bay, S.D. and Pazzani, M.J. (1999). Detecting change in categorical data: Mining contrast sets.

Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining, 302–306.

Bay, S.D. and Pazzani, M.J. (2001). Detecting group differences: Mining contrast sets. *Data Mining & Knowledge Discovery*, **5(3)**, 213–246.

Bayardo, R.J. (1998). Efficiently mining long patterns from databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data,* 85-93.

Boettcher, M. (2011). Contrast and change mining, *Data Mining & Knowledge Discovery*, **1**, 215–230.

Fisher, R.A. (1936). UCI repository of machine learning databases. URL https://archive.ics.uci.edu/ml/datasets/iris, Accessed on 21st June 2021.

Gupta, G. K. and Chauhan, G. S., (2005). Symptoms, identification and management of soybean diseases. Technical bulletin no. 10, National Research Centre for Soybean (ICAR), Indore, 92.

Han, J., and Kamber, M. (2006). Data mining: concepts and techniques, 2nd. *University of Illinois at Urbana Champaign: Morgan Kaufmann*.

Hartman, G.L., Sinclair, J. B. and Rupe, J.C., (1999). Compendium of Soybean Diseases, IV edition. The American Phytopathological Society. Academic press, St. Paul, Minnesota, 100.

Jain, R. and Arora, A. (2013). Approach for Mining Multiple Patterns from Clusters. *Journal of the Indian Society of Agricultural Statistics*, **67(1)**, 33-42.

Kaneiwa, K., and Kudo, Y. (2011). A sequential pattern mining algorithm using rough set theory. *International Journal of Approximate Reasoning*, **52(6)**, 881-893.

Kralj, P., Lavra, N., Gamberger, D. and Krsta, A. (2007). Contrast Set Mining through Subgroup Discovery Applied to Brain Ischaemia Data. *Advances in Knowledge Discovery and Data Mining,* **1,** 579-586.

Langohr, L., Podpecan, V., Petek, M., Mozetic, I., Gruden, K., Lavrac, N. and Toivonen, H. (2013). Contrasting Subgroup Discovery. *The Computer Journal*, **56(3)**, 289–303.

Magalhães, A. and Azevedo, P.J. (2014). Contrast set mining in temporal databases. *Expert Systems*, **32(3)**, 435–443.

Michalski, R.S. (1980). UCI repository of machine learning databases. URL https://archive.ics.uci.edu/ml/datasets/soyabean, Accessed on 12th February 2020.

Novak, P.K., Lavrac, N., Gamberger, D. and Krstacic, A. (2009). CSM-SD: Methodology for contrast set mining through subgroup discovery. *Journal of Biomedical Informatics,***42,** 113–122.

Novak, P.K., Lavrac, N., Webb, G.I. (2009). Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. *Journal of Machine Learning Research,***10**, 377-403.

Oreski, D. and Konecki, M. (2016). Handling Sparse Data Sets by Applying Contrast Set Mining in Feature Selection. *Journal of Software,***11(2)**, 148-161.

Qian, R., Yu, Y., Park, W., Murali, V., Fink, S., and Chandra, S. (2020). Debugging crashes using continuous contrast set mining. *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Software Engineering in Practice*, 61-70.

Rymon, R. (1992). Search through systematic set enumeration. *KR'92 Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, 539-550.

ANNEXURE I

Chi square table

Degrees of Freedom	Probability									
	Non Significant								Significant	Highly Significant
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21

Annexure II

Variable information of Soybean dataset

	Attribute : Attribute value
v1	date: april=0, may=1, june=2, july=3, august=4, september=5, october=6
v2	plant-stand: normal=0, lt-normal=1
v3	precip: lt-norm=0, norm=1, gt-norm=2
v4	temp: lt-norm=0, norm=1, gt-norm=2
v5	hail: yes=0, no=1
v6	crop-hist: diff-lst-year=0, same-lst-yr=1, same-lst-two-yrs=2, same-lst-sev-yrs=3
v7	area-damaged: scattered=0, low-areas=1, upper-areas=2, whole-field=3
v8	severity: pot-severe=1, severe=2
v9	seed-tmt: none=0, fungicide=1
v10	germination: '90-100%'=0, '80-89%'=1, 'lt-80%'=2
v12	leaves: norm=0, abnorm=1
v20	lodging: yes=0, no=1

	Attribute : Attribute value
v21	stem-cankers: absent=0, below-soil=1, above-soil=2, above-sec-nde=3
v22	canker-lesion: dna=0, brown=1, dk-brown-blk=2, tan=3
v23	fruiting-bodies: absent=0, present=1
v24	external decay: absent=0, firm-and-dry=1
v25	mycelium: absent=0, present=1
v26	int-discolor: none=0, black=2
v27	sclerotia: absent=0, present=1
v28	fruit-pods: norm=0, dna=3
v35	roots: norm=0, rotted=1

Annexure III

Discretized values of Iris dataset

Attribute	Original Value (in cms)	Discretized value
Sepal Length	infinite - 5.5	0
Sepal Length	5.5 - 6.7	1
Sepal Length	6.7 - infinite	2
Sepal Width	infinite - 2.8	0
Sepal Width	2.8 - 3.6	1
Sepal Width	3.6 - infinite	2
Petal Length	infinite - 2.967	0
Petal Length	2.967 - 4.933	1
Petal Length	4.933 - infinite	2
Petal Width	infinite - 0.9	0
Petal Width	0.9 - 1.7	1
Petal Width	1.7 - infinite	2