Nat. Hazards Earth Syst. Sci. Discuss., 3, 5957–5975, 2015
www.nat-hazards-earth-syst-sci-discuss.net/3/5957/2015/
doi:10.5194/nhessd-3-5957-2015

Natural Hazards and Earth System Sciences
Open Access
Discussions

Setting up the critical rainfall line for debris flows via support vector machines

Y. F. Tsai[1], **C. H. Chan**[2], **and C. H. Chang**[3]

[1]Department of Social and Regional Development, National Taipei University of Education, Taipei, Taiwan
[2]Department of Geography, National Taiwan University, Taipei, Taiwan
[3]Department of Civil Engineering, Tamkang University, New Taipei City, Taiwan

Received: 18 August 2015 – Accepted: 6 September 2015 – Published: 2 October 2015

Correspondence to: C. H. Chan (r01228005@ntu.edu.tw)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

The Chi-Chi earthquake in 1999 caused tremendous landslides which triggered many debris flows and resulted in significant loss of public lives and property. To prevent the disaster of debris flow, setting a critical rainfall line for each debris-flow stream is necessary. Firstly, 8 predisposing factors of debris flow were used to cluster 377 streams which have similar rainfall lines into 7 groups via the genetic algorithm. Then, support vector machines (SVM) were applied to setup the critical rainfall line for debris flows. SVM is a machine learning approach proposed based on statistical learning theory and has been widely used on pattern recognition and regression. This theory raises the generalized ability of learning mechanisms according to the minimum structural risk. Therefore, the advantage of using SVM can obtain results of minimized error rates without many training samples. Finally, the experimental results confirm that SVM method performs well in setting a critical rainfall line for each group of debris-flow streams.

1 Introduction

Taiwan is a mountainous island with very steep terrain and fragile geology. The extremely heavy rainfall caused by typhoon and Mei-Yu yearly often lead to large-scale debris flow disasters in mountains of Taiwan. Especially after the Chi-Chi earthquake in 1999, a lot of landslides have occurred in the center of Taiwan. These tremendous landslides often lead to the amount of sediment material on the streambed in the initiation area of debris flow. These sediment material will mobilize by the rainfall and cause numerous debris flow disasters which result in significant loss of public lives and property in the following rainy or typhoon seasons. Furthermore, the landslides triggered by the Chi-Chi Earthquake will have a significantly upward trend in scale and frequency. It means that the debris flow disasters have been more unpredictable and destructive with the amount of sediment material. As if the literature of Shieh et al. (2009), it showed that the numerous landslides triggered by Chi-Chi Earthquake caused a lot of debris-flows and lowered the rainfall threshold of these debris-flows in subsequent years. Nakamura et al. (2000) also indicated that there was a huge number of landslides for about 42 years after the Kanto earthquake in Japan, and almost every landslides during that time induced server debris-flow disasters, presented in Fig. 1.

Thus, in order to prevent the disaster of debris flow, setting a critical rainfall line for each debris-flow stream is necessary. In this research, we aims to setting a critical rainfall line for each debris-flow stream via a series of statistical method. Firstly, 377 debris-flow streams in the center of Taiwan affected by Chi-Chi earthquake was considered. Then, 8 predisposing factors of debris flow were used to cluster streams into 7 groups via the genetic algorithm. After streams with similar characteristics were clustered together, support vector machines (SVMs) were applied to setup the critical rainfall line for each debris-flow clusters. Finally, the experimental result shows that SVM method performs well in setting a critical rainfall line for each group of debris-flow streams.

2 Study area

The Chi-Chi earthquake occurred in 1999 in Taiwan caused numerous landslides, the blocks of these landslides were up to 2365 units and the area were approximate 14 347 ha, represented in Fig. 2a. These landslides mostly located in the mountains of central Taiwan. The debris flow streams triggered by the landslides which affected seriously by the Chi-Chi earthquake were studied in the experiments. 377 debris flow streams were chosen from 7 counties included Nan-Tou county, Maio-Li county, Taichung City, Taichung county, Chun-Chua county, Yun-Lin county and Chia-Yi county, represented in Fig. 2b.

According to Shieh and Tsai (2001), 8 important characteristics of the 377 debris-flow streams including rock type (R), watershed area (A), effective watershed area (A_{15}), landslide area (A_s), landslide ratio (A_s/A), length of channel in the effective watershed area (L), mean surface slope of the effective watershed area (S_s) and mean channel slope of the effective watershed area (S_c). Table 1 shows the eight characteristics of the 377 streams.

3 Data processing

In order to cluster the 377 debris flow streams into different groups, the statistical data including geographical information, hydrological information, historical data of disaster and statistical tables of eight predisposing factors have to preprocessed. The preprocessing involved presentation of the data, normalization of the data and the measurement of the distance between two debris-flow streams. Equation (1) represents the normalization of the data (Z-score).

$$Z_{ij} = \frac{x_{ij} - \overline{x_i}}{\sigma_i} \text{ where } 1 \le j \le K,\ 1 \le i \le M, \tag{1}$$

where M represents the 377 debris-flow streams and K is the eight attributes of each streams. Let F_i represent the ith debris flow where $1 \le j \le K$, and F_j represent the jth attribute of 8 predisposing factors where $1 \le i \le M$. The corresponding attribute vector for the debris flow F_i is represented as X_{ij}. The $\overline{X}$ and σ_i are the mean and mean absolute deviation of X_{ij}, respectively.

After the data were normalized, the distance between two debris-flows could be calculated. The centered Person correlation was used to define the distance $D(F_i, F_j)$. Let $F_i = (X_{i1}, X_{i2}, \ldots X_{ik})$ and $F_j = (X_{j1}, X_{j2}, \ldots X_{jk})$ be normalized attribute vectors of two flows over a series of K attributes. The distance between flows F_i and F_j was defined as Eq. (2).

$$S_{i,j} = \frac{1}{K} \sum_{l=1}^{k} \left[\frac{X_{il} - \overline{X}_i}{\sigma_{X_i}} \right] \left[\frac{X_{jl} - \overline{X}_j}{\sigma_{X_i}} \right], \text{ where } -1 \le S_{i,j} \le 1 \tag{2}$$

$$\overline{X_i} = \frac{\sum_{l=1}^{k} F_{il}}{k} \quad (3)$$

$$\sigma_{x_i} = \sqrt{\frac{1}{k}\sum_{l=1}^{k}\left(X_{il} - \overline{X}_i\right)^2} \quad (4)$$

Since this term measured distance, the following was defined as Eq. (5)

$$D\left(F_i, F_j\right) = 1 - S_{i,j}, \text{ where } -1 \leq D\left(F_i, F_j\right) \leq 1. \quad (5)$$

4 Clustering analysis of debris-flow steam

This section aims to cluster 377 debris-flow streams into seven groups via clustering analysis such that streams in each group have similar characteristics. An efficient clustering algorithm was considered for describing debris flows in order to illustrate the relationships by constructing a binary hierarchical tree. This approach was employed to group 377 debris flow streams into seven groups such that the critical rainfall line in the same group could be set. Many approaches to constructing binary hierarchical trees have been proposed. For example, Ward's method (Ward, 1963), the single-linkage method (Sibson, 1973), the average-linkage method (Defays, 1977), and the average-linkage (Voorhees, 1986) hierarchical clustering approach have been extensively applied in various fields to approximate such trees, including the fields of document clustering (Willet, 1988) and bioinformatics (Eisen et al., 1998; Alizadeh et al., 2000).

In this study, a means of family competition genetic algorithm (FCGA) was presented to construct a hierarchical tree of streams. The FCGA combines family competition, neighbor-join mutation (NJ) and edge assembly crossover (EAX) (Nagata and Kobayashi, 1997). The concepts of family competition have been successfully applied to solve numerous continuous parameter optimization problems, including protein docking (Yang and Kao, 2000). In the authors' earlier work (Tsai et al., 2001), family competition and EAX were successfully integrated to solve traveling salesman problems (TCPs). Neighbor-join mutation (Tsai et al., 2002) was developed to coordinate with the EAX and thus balance exploration and exploitation. The primary difference between the method in this study and that in our previous work is in the integration of these three mechanisms.

The experimental results revealed that the FCGA is a promising method for constructing the optimal tree of streams. Figure 3 presents the seven groups of 377 debris-flow streams. In Fig. 3a–g, the x axis represents the eight important characteristics and the y axis is the normalized values of each characteristics. Each groups include 39, 58, 61, 42, 67, 47 and 63 streams respectively. Each groups all exhibited different trends in their characteristics and the characteristics in the same group were similar. Additionally, it should be noted that Fig. 3f and g use different scales. The clustering results showed that the proposed method was able to cluster streams into separate groups with similar characteristics. As a result, this method represents a possible means of establishing a critical rainfall line for debris flow streams in each group. The critical rainfall lines of each groups could be set according to the characteristics.

5 Establishing the critical rainfall line for debris flows

When the streams with similar characteristics have clustered together, the critical rainfall line of debris flow could be set via SVM. SVM is a new machine learning approach proposed by Vapnic (1998) based on statistical learning theory. The advantage of SVM is that this theory raises the generalized ability of learning mechanisms according to minimize the structural risk. Therefore, we can obtain the results with minimum error rates and without many training samples. Otherwise, SVM is an optimized algorithm which can rapidly performed by a standard programming algorithm and obtained the global optima. The SVM has been widely applied in many disciplines to solve the problems of classification and regression in the field of hydrological engineering (Yu et al., 2011; Lin and Chen, 2011; Shen et al., 2011). This study tried to establish the critical rainfall line of debris flow via SVM.

In this study, each data of debris flow stream were consider as a vector or a point in a multidimensional space. The destination of this research is to find a hyper-plane which can separate the vectors into two parts, according to the occurrence of debris flow, represented in Fig. 4.

However, two problems have encountered frequently in most cases during the process of classification. Figure 5 shows the two problems, it is likely that there are many hyper-planes existed in the multidimensional space, or it does not have any hyper-plane could separate the training data into two parts exactly.

Therefore, we switched these training data to a higher dimensional space called feature space via a non-linear function $\varphi(x)$. In the feature space, we could still find several hyper-planes can separate the training data into two groups successfully. However, the method of SVM we applied can choose a particular plane from those hyper-planes named maximum margin hyper-plane. Figure 6 shows that with the advantage of SVM, a maximum margin hyper-plane could be selected from several hyper-planes. This maximum margin hyper-plane represents the inner product of the vector in the feature space. This inner product generally made by a kernel function, hence we can easily find the maximum margin hyper-plane with a suitable kernel function. Under the situation of maximum margin hyper-plane, the sum of distance from those training data closest to the plane would be maximum. The decision-making hyper-plane for classification could be illustrated with less training data near the hyper-plane called support vector. Therefore, these training data can classified easily and efficiently.

This section describes the result of the proposed method SVM to establish the critical rainfall line for each group of debris flows. When the debris flow streams with similar characteristics have clustered together into seven group, the critical rainfall line of each debris-flow group could be set via SVM. Figure 7a–g shows the critical rainfall line of group A, group B, group C, group D, group E, group F and group G, respectively. In Fig. 7, the blue dots represent the rainfall data with debris flow events and the purple dots symbolize the rainfall data without debris-flow event. The green zone represents a range of hazardous area with debris flow. In contrast, the black zone represents a range without debris flow. The boundary between the green area and black region stands for a critical rainfall line of each debris flow groups.

Compared with the critical rainfall lines of each groups respectively, it could find that the critical rainfall lines of group D and group F were lower than others, showed in Table 2. It is likely that the group F have a higher landslides ratio (A_s/A) and group D have a vulnerable rock type (R) and steep slope on mean channel of the effective watershed area (S_c) and mean surface of the effective watershed area (S_s) (Fig. 3). In contrast, group A and group E have higher critical rainfall lines with each characteristic lower than the average of other groups. As a result, we can establish the critical rainfall lines of each debris flow groups clustered together with similar characteristic via support vector machine. The critical rainfall lines were set according to the characteristic of debris flow.

6 Conclusions

This study aims to set up the critical rainfall line of debris flows via a series of statistical methods. 377 debris-flow streams in the center of Taiwan affected by Chi-Chi earthquake and 8 predisposing factors of debris flow were considered. Firstly, 377 debris flow streams were clustered into 7 groups with similar characteristic via family competition genetic algorithm, then support vector machine was used to set up the critical rainfall line of each debris flow groups. The results reveal that SVM can establish the critical rainfall lines of debris flow successfully and the critical rainfall lines were set according to the characteristic of each debris flow groups. Hence, the method proposed in this study can be an effective instrument for establishing critical rainfall lines. In the future, the weights and the interactions of the predisposing characteristics would be the focus of research.

Acknowledgements. This work was supported by the National Science Council of Taiwan (Grant No.: NSC-94-2211-E-152-001 and NSC 93-2211-E-152-001).

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson Jr., J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature, 403, 503–511, 2000.

Defays, D.: An efficient algorithm for a complete link method, Computer J., 20, 364–366, 1977.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D.: Cluster analysis and display of genome-wide expression patterns, P. Natl. Acad. Sci., 95, 14863–14868, 1998.

Lin, G. F. and Chen, L. H.: Typhoon characteristics and their effects on ann-based hourly flood forecasting models, in: The 2nd International Workshop on Multimodal Sediment Disaster-asian Cloud Network on Disaster Research, Tainan, Taiwan, 11–18, 2011.

Nagata, Y. and Kobayashi, S.: An analysis of edge assembly crossover for the traveling salesman problem, in: The Systems, Man, and Cybernetics, 1999, IEEE SMC '99 Conference, Tokyo, Japan, 1999.

Shen, C. W., Lo, W. C., and Chen, C. Y.: Using support vector machine to establish models to determine potential debris-flow creeks, J. Soil Water Conserv., 43, 203–218, 2011.

Shieh, C. L. and Tsai, Y. F.: An evaluation research on alert parts of the debris flow and the happening standard value, Soil and Water Conservation Bureau, Nantou, Taiwan, 37–49, 2001.

Shieh, C. L., Chen, Y. S., Tsai, Y. J., and Wu, J. H.: Variability in Rainfall Threshold for Debris Flow after the Chi-Chi Earthquake in Central Taiwan, China, Int. J. Sediment Res., 24, 177–188, 2009.

Sibson, R.: SLINK: An optimally efficient algorithm for the single-link cluster method, Computer J., 16, 30–34, 1973.

Tsai, Y. F. and Shieh, C. L.: Hazard zone mapping of debris flow, J. Mechanics, 18, 91–102, 2002.

Vapnik, V.: Statistical Learning Theory, John Wiley, New York, 1998.

Voorhees, E. M.: Implementing agglomerative hierarchic clustering algorithms for use in document retrieval, Inform. Process. Manage., 22, 465–476, 1986.

Ward, J. H.: Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc., 58, 236–244, 1963.

Willett, P.: Recent trends in hierarchic document clustering: A Critical Review, Inform. Process. Manage., 24, 577–597, 1988.

Yang, J. M. and Kao, C. Y.: Flexible ligand docking using a robust evolutionary algorithm, J. Comput. Chem., 21, 988–998, 2000.

Yu, P. S., Yang, T. C., Ku, C. M., and Lin, M. X.: Comparing support vector machine and fuzzy inference model for real-time updating of flow forecasting, in: 2011 International Symposium on Natural Disaster Prevention-coping with Extreme Flood and Sediment-related Disasters, Tainan, Taiwan, 35–43, 2011.

Table 1. The statistical table of 8 characteristics with 377 streams.

R		A_s/A (%)		S_c (degree)		S_s (degree)	
Type	Stream no.	Range	Stream no.	Range	Stream no.	Range	Stream no.
Alluvion	4	0–5	3	0–10	3	0–0.5	81
Conglomerate	30	5–10	9	10–15	11	0.5–1.0	35
Sand stone	35	10–15	67	15–20	42	1.0–2.0	65
Sand and shale	191	15–20	115	20–25	80	2.0–3.0	39
Shale	5	20–25	99	25–30	103	3.0–5.0	45
Slate	55	25–30	56	30–35	93	5.0–10.0	49
Metamorphic sand	57	> 30	28	> 35	45	> 10.0	63

A (ha)		A_{15} (ha)		A_s (ha)		L (km)	
Range	Stream no.	Range	Stream no.	Range	Stream no.	Range	Stream no.
0–10	13	0–10	23	0–0.1	59	0–0.5	54
10–20	19	10–20	34	0.1–0.5	45	0.5–1.0	88
20–30	18	20–30	22	0.5–1.0	37	1.0–1.5	78
30–40	32	30–40	30	1.0–2.0	53	1.5–2.0	44
40–50	17	40–50	28	2.0–3.0	25	2.0–3.0	53
50–60	19	50–60	23	3.0–4.0	30	3.0–4.0	19
60–70	18	60–70	18	4.0–5.0	23	4.0–5.0	10
70–80	13	70–80	13	5.0–10.0	38	5.0–6.0	11
80–90	19	80–90	19	10.0–20.0	29	6.0–7.0	4
90–100	14	90–100	12	20.0–30.0	15	7.0–8.0	4
100–200	91	100–200	71	30.0–40.0	9	8.0–9.0	1
200–300	30	200–300	27	40.0–100.0	8	9.0–10.0	4
300–500	36	300–500	25	> 100.0	6	> 10.0	7
> 500	38	> 500	32				

Table 2. The value of critical rainfall line of each debris flow groups.

Group	Group A	Group B	Group C	Group D	Group E	Group F	Group G
Rainfall intensity	31.5	23	18	23.5	40.5	22.5	18.5
Rainfall accumulation	235.5	195	195	70	218	50	280

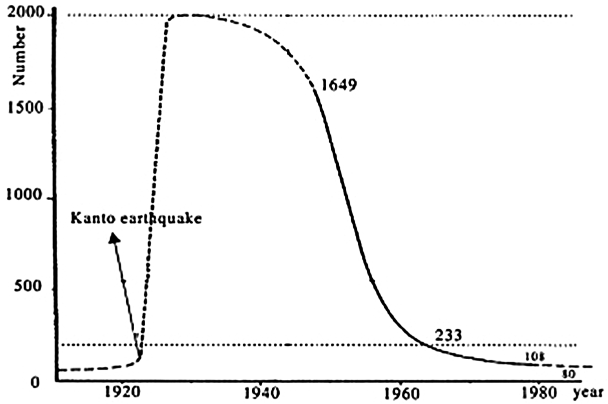

Figure 1. The statistical information of the landslides after the Kanto earthquake in Japan (Nakamura et al., 2000).

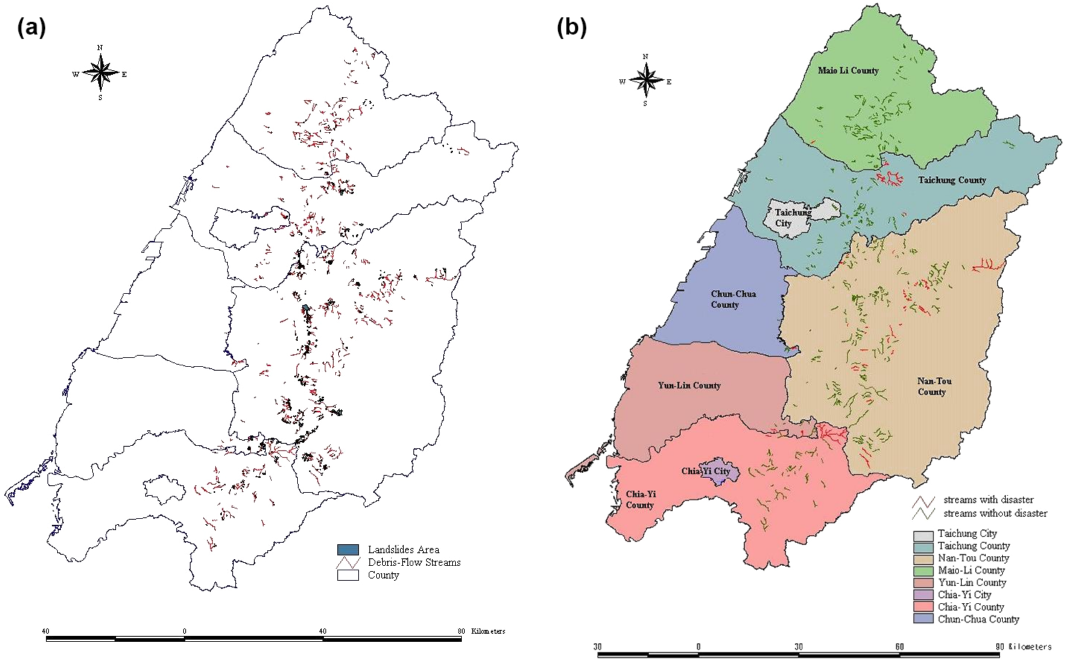

Figure 2. (a) The position and landslide area of 377 debris-flow streams in central Taiwan and **(b)** he historical disaster and position of 377 debris-flow streams in central Taiwan.

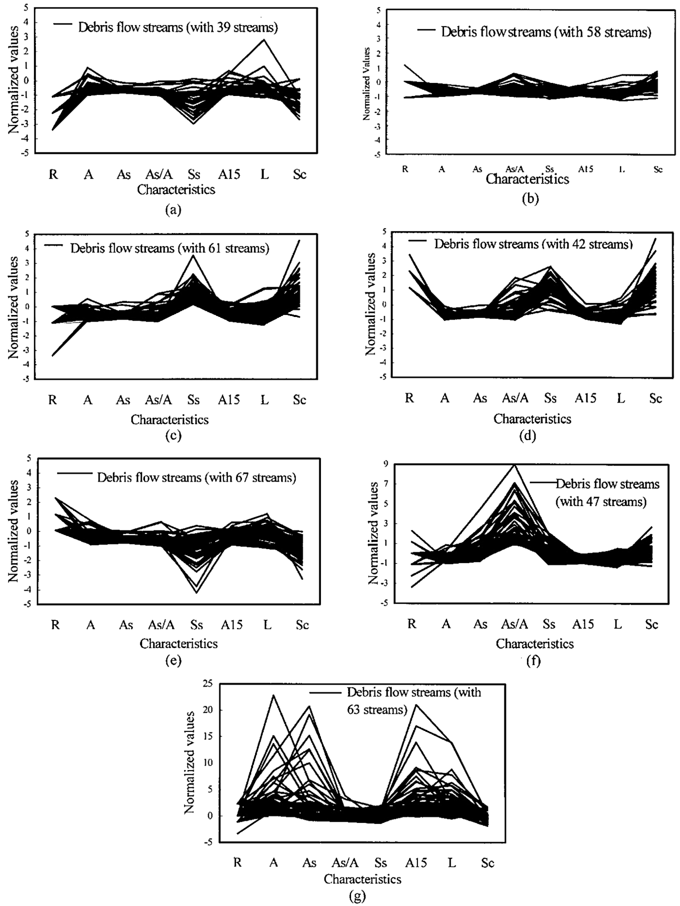

Figure 3. The results of clustering analysis on 377 debris-flow streams.

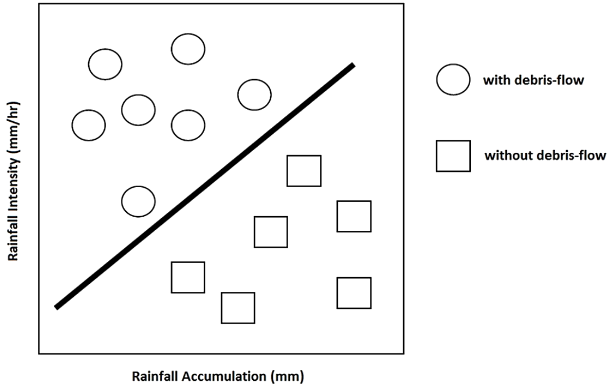

Figure 4. The multidimensional space with vector of debris-flow.

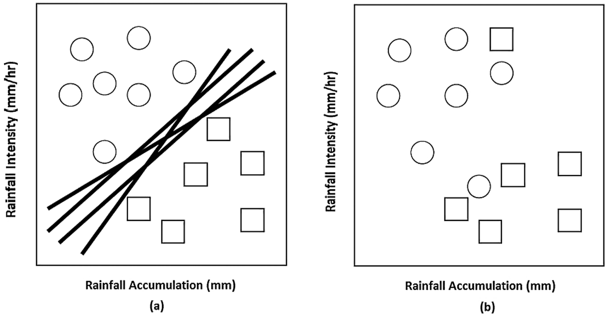

Figure 5. The examples of problems encountered in most cases.

NHESSD
3, 5957–5975, 2015

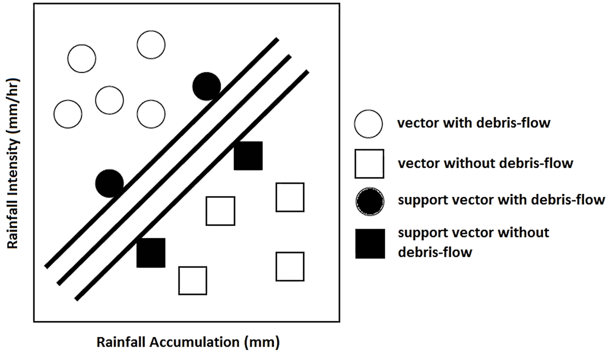

Figure 6. An example of maximum margin hyper-plane and support vector.

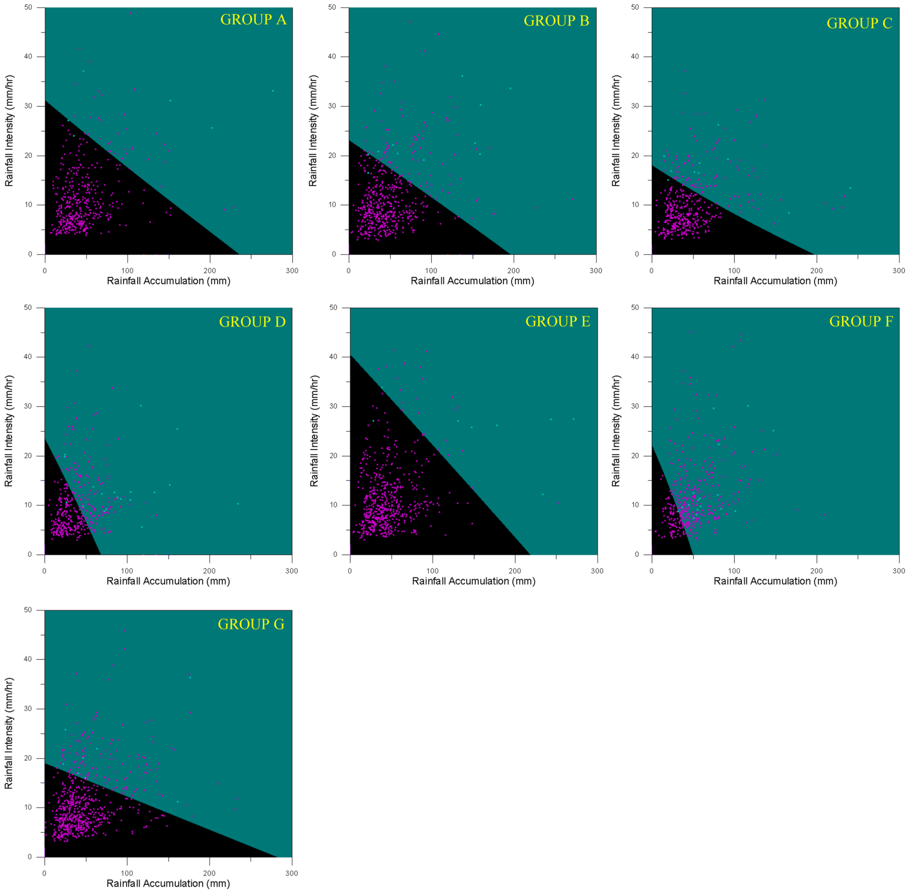

Figure 7. The result of our research tested on 7 groups.