# Adapting cosine similarity for avalanche forecasting using NN model

Neha Ajit Kushe[1], Dr. Ganesh M. Magar[2]
[1]Research Scholar, [2]Associate Professor
*P. G. Department of Computer Science,*
*S.N.D.T. Women's University, Mumbai – 49*
*Email: neha.kushe@hotmail.com*

***Abstract-*** Various methodologies of pattern matching have been used for forecasting of snow avalanche. Nearest neighbour was used to find the similar days with regards to the present day and forecast the avalanche situation for the present day based on the avalanche conditions for all the neighbouring days. Euclidean distance was used as a metric for finding the distance between the present day and past days. However instead of distance, the results of using similarity have been encouraging in pattern matching. Recent study has been found that used cosine similarity with nearest neighbour algorithm for avalanche forecasting. Present work focusses on developing a new version of the said model by also considering the change in the values of the parameter over 24 hours along with raw variables. This study has been conducted by using the snow and meteorological data of Bahang region of Himachal Pradesh in India. The forecasted results were compared with the observed avalanche situations and the accuracy of the model in terms of performance measures were obtained.

***Keywords****:* Avalanche; Bahang; Cosine Similarity; Forecasting; Nearest Neighbour

## I. INTRODUCTION

Use of nearest neighbour for finding similar days with regards to the present day has been observed in various avalanche prediction models. All these models are based on the concept that if an avalanche had occurred during a particular weather condition, then presence of similar weather conditions can lead to an avalanche. Usually Euclidean distance has been used as a distance metric for calculating the distance between the days and to find ten most similar days[1]–[5]. Depending on the avalanche condition on these ten days, it is decided whether the present day is an avalanche day. Application of nearest neighbour to find probability of avalanche was put forth by [6]. Initial attempts to use nearest neighbour for forecasting of avalanche were by [2], [7], [8]. Machine learning functions were used to improve the reliability of avalanche prediction by combining artificial intelligence techniques with NXD[9]. Model was suggested by taking variables at an interval 3 hrs instead of 24 hrs for avalanche forecasting model[10]. A model was put forth by [11] that used clustering techniques for the analysis of the nearest neighbours by using Mahalanobis as a distance metric between data vectors. Modifications were undertaken to the initial model to consider the application of explanatory or elaborate variable by [3]. Model was made to provide avalanche forecasts at a regional level by [1]. Other variants of nearest neighbour models were found at Scotland, France, British Columbia, USA and India[4], [5], [12]–[18].

According to [2], "We are free to define the distance between two days (n-tuples) in whatever way suits our purpose and we choose the most common one: $r^2 = \sum_i \Delta x_i^2$". Majority of the avalanche forecasting models have used Euclidean distance as a distance metric. The basic working of Euclidean distance as a distance metric is that it calculates the distance between the present day and the historical days and those days which have the least Euclidean distance are considered as the closest day. A variant to the nearest neighbour model for avalanche forecasting has been put forth by [19] which uses the cosine similarity to find the closest day to the present day. This model considered only the raw data that was obtained from the station. Changes in weather condition due to time change were not included in the model. Present study uses cosine similarity with nearest neighbour model to find avalanche occurrence by taking into consideration the change in the values of parameters over a period of 24 hrs along with raw variables.

## II. INFLUENCE OF PREVIOUS DAY IN AVALANCHE FORECASTING

In [19], the behaviour of the parameters was only considered by using raw values of parameters. The influence of the past weather conditions was not taken into consideration while calculating the neighbours of a particular day. For our current research, the parameters selected for the model are the combination of the raw variables and the parameters over a period of 24 hrs. Considering the parameters over 24 hrs help in forecasting as avalanches can also occur as the effect of change in the parameters over a period of time. For experimental purpose, the effect of snowfall over 24 hrs, 48 hrs and 72 hrs were checked. However, for the region under study, consideration of change over 48 hrs and 72 hrs did not have any positive effect on the output of the model. The major reason for this was that the region used for study does not have high frequency of avalanches and as a result, consecutive days may not have similar weather conditions. Because of this, only the change in parameters over a period of 24 hrs was used for the current study

## III. STUDY AREA AND DATA CHARACTERISTICS

The present study has been conducted by using the snow and meteorological data of Bahang region in Himachal Pradesh. The area belongs to the Pir Panjal range which falls in the Lower Himalayan zone or the Sub tropical zone[20]. The frequency of avalanche activity for this region is low. Snow & Avalanche Study Establishment (SASE) had installed an observation station at Bahang (2192 m a.s.l) for obtaining the daily snow and weather data[21]. The rough geographical extent of this region is shown in Figure 1.

The past records for 22 snow meteorological variables from year 2005 to 2013 for this region were archived in the database. These records were taken for the months of January to March. The Avalanche information for all these days was also stored in the database. The records for all the days were taken at 0830 hrs and 1730 hrs (both IST (GMT + 5.30)). The relation between humidity, wet temperature and dry temperature as defined by [22] was used to find the missing values. The values that were not acquired even after this process were distinctively marked. Total records of 812 days were used for the present study. Out of these 22 variables, a total of 9 variables were selected for the proposed work. These were selected based on their variation with regards to avalanche occurrence and are a combination of raw variables obtained from the weather station installed at Bahang and derived variables. The list of variables that were used for our current study are given in Table 1.

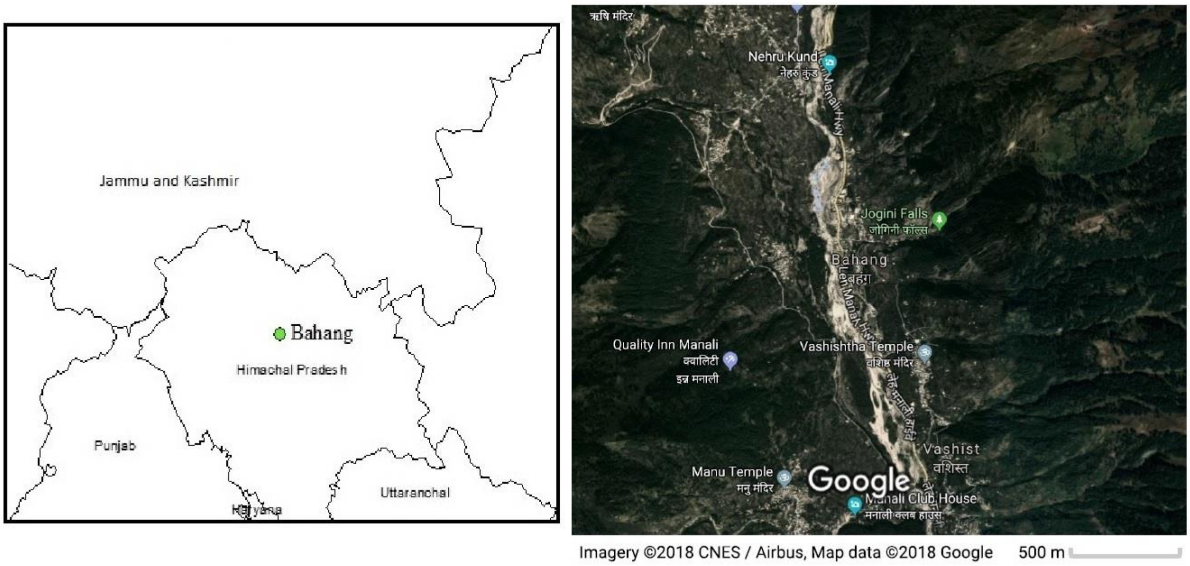


Figure 1: Study Area – Bahang (32.16.33, 77.09.03)

Table 1: Snow Meteorological Variables Considered for present study

| Sr. No | Variable | Day/Time | Unit |
|---|---|---|---|
| 1 | Wet Temperature | 0830 hrs (t) | ºC |
| 2 | Humidity in 24 hrs | 0830(t-1)-0830 hrs (t) | % |
| 3 | Average Wind Speed in 24 hrs | 0830(t-1)-0830 hrs (t) | kmph |
| 4 | Snow Height | 0830 hrs (t) | cm |
| 5 | Snow Temperature | 0830 hrs (t) | ºC |
| 6 | Fresh Snow in 24 hrs | 0830(t-1)-0830 hrs (t) | cm |
| 7 | Snow Water Equivalent in 24 hrs | 0830(t-1)-0830 hrs (t) | mm |
| 8 | Rate of Snow Fall in 24 hrs | 0830(t-1)-0830 hrs (t) | cm/min |
| 9 | Snow Penetration | 0830 hrs (t) | cm |

## IV. FORECASTING ALGORITHM

The proposed model finds the similar days by using cosine similarity to find the closeness between the days in combination with the nearest neighbour algorithm. The methodology applied in the proposed model is described in [19]. Cosine Similarity represents the cosine of the angle between the two vectors which has been very useful to find similar patterns in a dataset. If A and B are two vectors and $A_i$ and $B_i$ are its components respectively, then Cosine Similarity is given by

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

The steps used for finding the avalanche forecast are as follows:

1. Compute the Cosine Similarity between the present day and the each of the past days using equation (1).
2. Arrange the past days depending on their cosine similarity value in the decreasing order i.e. higher the cosine similarity, closer is the past day to present day.
3. Obtain the nearest neighbours to the present days by taking first 10 days from the order obtained in step 2.
4. Find the avalanche occurrence situation for all the ten neighbours.
5. Probability of avalanche occurrence p(A) is found out by
   $$p(A) = \frac{nA}{N} \quad (2)$$
   where nA is the number of avalanche days present in the nearest neighbours and N is the total number of neighbours considered.
6. If the probability of avalanche occurrence is greater than or equal to threshold value, then the present day is considered as an avalanche day.

For our model, number of nearest neighbours considered are ten [8] and the threshold value is considered as 0.3 [7]. Depending on the present day, the output of our model can be 0 (Non-avalanche Day) or 1(Avalanche day). This type of event is called as binary event [23]. The number of possible outcomes in case of binary events are given in Table 2.

Table 2: Possible Outcomes for our model

| Forecasted \ Observed | Yes | No |
|---|---|---|
| Yes | True Positive (a) | False positive (b) |
| No | False Negative (c) | True Negative (d) |
| Total Outcomes (n) =a+b+c+d | | |

For the current study, the dataset is divided into training dataset and testing dataset. The model is trained for the training dataset and then the working of the model is observed by applying the model on the testing dataset. For our study purpose, the model was applied on three combinations of training and testing dataset. The application of the model on the various combinations of the testing and training dataset helps to understand the output of the model for different situations. The combinations of dataset are summarized in Table 3.

Table 3: Combinations for Training and Testing Dataset

| Dataset No. | Training Dataset | Testing Dataset |
|---|---|---|
| I | Jan 2005 to March 2010 | Jan 2011 to March 2013 |
| II | Jan 2005 to March 2011 | Jan 2012 to March 2013 |
| III | Jan 2005 to March 2012 | Jan 2013 to March 2013 |

## V. PERFORMANCE MEASURES USED FOR THE CURRENT STUDY

The results for all three combinations of training and testing dataset were evaluated by using performance measures[23]. These are the verification measure that focus on the correspondence between the observations and the forecasts. The performance measures used for this study are: Heidke Skill Score (HSS) and Proportion Correct (PC).

Heidke Skill Score (HSS) is a skill score used frequently for summarizing squared contingency table[24]. HSS provides the fraction of the correct forecast after eliminating the forecast that have been correct by chance. HSS are better representative of forecast accuracy when the cases of "YES" events are comparatively less than "NO" events throughout the dataset[17]. HSS is defined as

$$\text{HSS} = \frac{\text{ad}-\text{bc}}{\frac{(\text{a+c})(\text{c+d})+(\text{a+b})(\text{b+d})}{2}} \quad (3)$$

Proportion Correct (PC) provides the proportion of the data sample that is a correct forecast for either kind i.e. 0 or 1 [12]. Hence is the ratio of summation of the proper forecasted YES events and proper forecasted NO events with regards to the total sample. It is defined by

$$PC = \frac{a+d}{n} \quad (4)$$

## VI. RESULTS

Nearest Neighbour Implementation using cosine similarity was implemented using the parameters mentioned in Table 1. The output for all the three combinations of the dataset was obtained. From the results, it is observed that for all the dataset combinations, the avalanche occurrence values are obtained properly by our model. This is seen as value for false negatives for all the dataset combinations is zero. However, a considerable number of false positives are observed in case of all the datasets. This implies that the model has misclassified certain days to be avalanche days whereas in reality they were non – avalanche days. These values are also further reflected in our performance measures which provides the correctness of the model in terms of different metrics. Table 4 contains the performance measures values for all the combinations of training and testing dataset.

Table 4: Consolidated Results for the model

| Data set No. | PC | HSS |
|---|---|---|
| I | 0.95572 | 0.385488 |
| II | 0.966851 | 0.487252 |
| III | 0.988889 | 0.851485 |

For both the performance measures, the value for perfect forecast is 1. The values of Proportion Correct (PC) are encouraging as the model is able to provide correct forecast for more than 90% of the days that are present in the dataset. However, it can be seen from the HSS values, the working of the model in case of dataset I and dataset II is far from a perfect forecast. For dataset III, the value of HSS is encouraging.

## VII. CONCLUSION:

Avalanche forecasting model that took into consideration the change in the values of parameters over a period of 24 hrs based on nearest neighbour algorithm along with cosine similarity was implemented. The accuracy of the model with regards to the performance measures was found to be encouraging. Compared to the model developed with raw variables, this model is providing better results. However, proper tuning of the model is required to eliminate the false positive values and provide better forecast.

## ACKNOWLEDGMENT

The Authors are thankful to Director, SASE (Snow Avalanche Study Establishment) for providing us with the snow and meteorological data for Bahang, Himachal Pradesh for our research study purpose.

## REFERENCES

[1] Bernhard Brabec and Roland Meister, "A Nearest Neighbor Model For Regional Avalanche Forecasting," *Ann. Glaciol.*, vol. 32, pp. 130–134, 2001.

[2] O. Buser, "Avalanche forecast with the method of nearest neighbours: an interactive approach," *Cold Reg. Sci. Technol.*, vol. 8, no. 2, pp. 155–163, 1983.

[3] Martin Gassner, Hans-Jurg Etter, Karl Birkeland, and Tom Leonard, "Nxd2000: An Improved Avalanche Forecasting Program Based On The Nearest Neighbor Method," in *Proceedings Of The International Snow Science Workshop*, Big Sky, Montana, USA, 2000, pp. 52–59.

[4] L. Mérindol, G. Guyomarc'h, and G. Giraud, "A French Local Tool For Avalanche Hazard Forecasting: Astral, Current State And New Developments," in *Proceedings of the International Snow Science Workshop, Pentiction, Canada*, 2002, pp. 105–108.

[5] Ross Purves, Keith Morrison, Graham Moss, and Blyth Wright, "Cornice- Development Of A Nearest Neighbours Model Applied In Backcountry Avalanche Forecasting In Scotland," in *Stevens, J.R. (Ed.), Proceedings Of International Snow Science Workshop*, Penticton, B.C., Canada, 2002, pp. 117–122.

[6] Charles Obled and Walter Good, "Recent Developments of Avalanche Forecasting by Discriminant Analysis Techniques: A Methodological Review and Some Applications to the Parsenn Area (Davos, Switzerland)," *J. Glaciol.*, vol. 25, no. 92, pp. 315–346, 1980.

[7] O. Buser and W. Good, "Avalanche Forecast: Experience Using Nearest Neighbors," in *Int. Snow Science Workshop, Aspen, Col*, Aspen, Colardo, 1984, pp. 109–115.

[8] Othmar Buser, Monika Butler, and Walter Good, "Avalanche forecast by the nearest neighbour method," *Int. Assoc. Hydrol. Sci. Publ.*, vol. 162, pp. 557–569, 1987.

[9] R. Bolognesi, O. Buser, and W. Good, "Local avalanche forecasting in Switzerland: strategy and tools, a new approach," in *International Snow Science Workshop*, 1994, pp. 463–472.

[10] K. Kristensen and C. Larsson, "An Avalanche Forecasting Program Based On A Modified Nearest Neighbour Method," in *Proceedings of the 1994 International Snow Science Workshop*, Snowbird, Utah, USA, 1994, pp. 22–30.

[11] D. M. McClung and J. Tweedy, "Numerical avalanche prediction: Kootenay Pass, British Columbia, Canada," *J. Glaciol.*, vol. 40, no. 135, pp. 350–358, 1994.

[12] P. Cordy, D. M. McClung, C. J. Hawkins, J. Tweedy, and T. Weick, "Computer assisted avalanche prediction using electronic weather sensor data," *Cold Reg. Sci. Technol.*, vol. 59, no. 2–3, pp. 227–233, Nov. 2009.

[13] C. McCollister, K. Birkeland, K. Hansen, R. Aspinall, and R. Comey, "Exploring multi-scale spatial patterns in historical avalanche data, Jackson Hole Mountain Resort, Wyoming," *Cold Reg. Sci. Technol.*, vol. 37, no. 3, pp. 299–313, Nov. 2003.

[14] C. McCollister, K. Birkeland, K. Hansen, R. Aspinall, and R. Comey, "A probabilistic technique for exploring multi-scale spatial patterns in historical avalanche data by combining GIS and meteorological nearest neighbors with an example from the Jackson Hole Ski Area, Wyoming," in *Proceedings of the 2002 International Snow Science Workshop*, Penticton, B.C., Canada, 2002, pp. 109–116.

[15] R. . Purves, K. . Morrison, G. Moss, and D. S. . Wright, "Nearest neighbours for avalanche forecasting in Scotland—development, verification and optimisation of a model," *Cold Reg. Sci. Technol.*, vol. 37, no. 3, pp. 343–355, Nov. 2003.

[16] A. Singh, K. Deep, and P. Grover, "A novel approach to accelerate calibration process of a k -nearest neighbours classifier using GPU," *J. Parallel Distrib. Comput.*, vol. 104, pp. 114–129, Jun. 2017.

[17] A. Singh, B. Damir, K. Deep, and A. Ganju, "Calibration of nearest neighbors model for avalanche forecasting," *Cold Reg. Sci. Technol.*, vol. 109, pp. 33–42, Jan. 2015.

[18] A. Singh and A. Ganju, "A supplement to nearest-neighbour method for avalanche forecasting," *Cold Reg. Sci. Technol.*, vol. 39, no. 2–3, pp. 105–113, Oct. 2004.

[19] Neha Ajit Kushe and Dr. Ganesh M. Magar, "An alternative approach with nearest neighbour classifier for forecasting snow avalanche," *Int. J. Comput. Eng. Appl.*, vol. 12, no. 2, Feb. 2018.

[20] S. S. Sharma and A. Ganju, "Complexities of avalanche forecasting in Western Himalaya—an overview," *Cold Reg. Sci. Technol.*, vol. 31, no. 2, pp. 95–102, 2000.

[21] H. S. Gusain, V. D. Mishra, and M. R. Bhutiyani, "Winter temperature and snowfall trends in the cryospheric region of north-west Himalaya," *Mausam India*, vol. 65, no. 3, pp. 425–432, 2014.

[22] R. Stull, "Wet-Bulb Temperature from Relative Humidity and Air Temperature," *J. Appl. Meteorol. Climatol.*, vol. 50, no. 11, pp. 2267–2269, Nov. 2011.

[23] I. T. Jolliffe and D. B. Stephenson, Eds., *Forecast verification: a practitioner's guide in atmospheric science*. Chichester: Wiley, 2003.

[24] D. S. Wilks, *Statistical methods in the atmospheric sciences*, 2. ed., [Nachdr.]. Amsterdam: Elsevier [u.a.], 2009.