Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 6; October-December, 2016, pp. 502-505
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

A Dissimilarity Margin Based Feature Selection Method for Interval Data

Irani Hazarika[1], Ragini Mishra[2] and Anjana Kakoti Mahanta[3]

[1,2,3]Dept of Computer Science Gauhati University
E-mail: [1]queensarathi@gmail.com, [2]ragini.mishra@gmail.com, [3]anjanagu@yahoo.co.in

Abstract—*In this work, we propose a feature selection method for interval data, which is based on standard feature selection algorithm Relief. In Relief algorithm a dissimilarity margin is defined for each data based on its nearest neighbor. But, finding nearest neighbor is a computationally complex task if the neighborhood is large. So, instead of nearest neighbor in this work we use the class prototype to define dissimilarity margin for the data, which will reduce the computational complexity of Relief algorithm. In experimental results, we compare the performance of our method with nearest neighbor based Relief algorithm using different dissimilarity measures for interval data.*

1. INTRODUCTION

A feature is an individual measurable property of the process being observed. There are many applications in machine learning or pattern recognition where data have huge numbers of features, many of which can be irrelevant or redundant. The focus of feature selection is to select a subset of features from these huge numbers of features by removing irrelevant features. Advantages of feature Selection (variable elimination) are- it helps in understanding data, reducing computation requirement, and improving the predictor performance. So, before applying any data mining techniques like clustering, classification on data, we can used the feature selection method as processing task on the data to improve the accuracy of clustering or classification.

There are various feature selection algorithms [1] found for different types of data such as, numeric, categorical, interval data etc which can be mainly divided into two types- filter and wrapper.

Filter methods

Filter feature selection methods apply a statistical measure called variable ranking technique to assign a scoring to each feature. Thus the features are ranked by the scores and a threshold is used to remove variables below the threshold. This method considers the feature independently, or with regard to the dependent variable.

Wrapper methods

Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model is used to evaluate a combination of features and assign a score based on model accuracy. The simplified algorithms such as sequential search or evolutionary algorithms such as Genetic Algorithm (GA) or Particle Swarm Optimization (PSO) are used in the predictive model.

One way in variable ranking methods is to compute the margin. When margin is calculated based of distances then it is called dissimilarity margin, otherwise it is called similarity margin. The dissimilarity margin is defined [2] as "The difference between the maximal distance for the false classes and the distance of the true class" and the similarity margin is defined as "The difference between the similarity for the true class and the maximal similarity of the false classes". In [3], the authors have proposed a margin based feature selection method and apply it to measure the quality of sets of features. The algorithm holds a weight vector over all features and updates this vector according to the sample points presented. In [4], the author proposed a method called Relief and it is proved that under some assumptions, the expected weight is large for relevant features and small for irrelevant ones. They also explain how to choose the relevance threshold τ in a way that ensures the probability that a given irrelevant feature will be chosen is small. Relief was extended to deal with multi-class problems, noise and missing data by [5].

In real world data may be found in the form of intervals. In interval data each features contains an interval of the form [low, high], where 'low' denotes the lower bound and 'high' denotes the upper bound. Feature selection for interval data is also an important issue. In [6] a similarity margin based feature selection method has been proposed for interval data, which uses similarity measures to compute closeness between two intervals. But there are many dissimilarity measures available for interval data. So, in this paper we have proposed a dissimilarity margin based feature selection method for interval data, which is based on Relief algorithm.

In section 2 related works, in section 3 preliminary definitions, in section 4 proposed methods, in section 5 experimental results, in section 6 conclusion has been given.

2. PRELIMINARIES

Interval

Given a totally ordered domain D, a non-empty subset I of D is called an interval iff for all *l*, r ϵ I and c ϵ D, $l \leq c \leq r$ implies c ϵ I. If for a given interval I, $l \leq c \leq r$ holds for all c ϵ D where *l* and r are two specific elements in I then I is denoted by [*l*, r], and *l* is the left end point (or lower bound) and r is called the right end point (or upper bound) of I. The interval I is a closed interval because for any c ϵ I, $l \leq c \leq r$.

Interval Pattern

An interval pattern I = {I_1, I_2,…., I_m} consists of 'm' interval variables such that each I_j ($1 \leq j \leq m$) contains an interval like [l, r] for I.

Distance Measures for Interval Data

There are various distance measures available in literature for interval data [7]. The distance measures which are based on L_p norm (for p=1, 2, ∞) are-

If $I_{xi}=[I^-_{xi}, I^+_{xi}]$, $I_{yi}=[I^-_{yi}, I^+_{yi}]$ are two intervals then distance between I_{xi} and I_{yi} can be defined as-

i. L_1-Norm: $d_{L_1}(I_{xi}, I_{yi})= |I^-_{xi} - I^-_{yi}|+ |I^+_{xi} - I^+_{yi}|$

ii. L_2-Norm: $d_{L_2}(I_{xi}, I_{yi})=\sqrt{|I^-_{xi} - I^-_{yi}|^2 + |I^+_{xi} - I^+_{yi}|^2}$

iii. L_∞-Norm: $d_{L_\infty}(I_{xi}, I_{yi})=\max (|I^-_{xi} - I^-_{yi}|, |I^+_{xi} - I^+_{yi}|)$

Again, if I_x={I_{x1}, I_{x2},.. I_{xm}} and I_y={I_{y1}, I_{y2},.. I_{ym}}, where $I_{xi}=[I^-_{xi}, I^+_{xi}]$, $I_{yi}=[I^-_{yi}, I^+_{yi}]$ denote i^th intervals of I_x, I_y respectively then distance between I_x and I_y is defined as-

$$D(I_x, I_y)= \sum_{i=1}^{m} d_{\text{dist}}(I_{xi}, I_{yi}).$$

Class Prototype for interval data

The class prototype must represent appropriately a group of samples, characterized by a vector of interval features and belonging to the same class. Let us consider a class c having N_c samples then each component of this vector is given by the following arithmetic means:

$$\rho_c^{j-} = \frac{1}{N_c}\sum_{i=1}^{N_c} x_i^{j-} \text{ and } \rho_c^{j+} = \frac{1}{N_c}\sum_{i=1}^{N_c} x_i^{j+}$$

Where x_i^{j-} is the jth feature lower bound of the i^th sample and x_i^{j+} is its upper bound. Therefore, each j^th component for the class c is then represented by an interval like

$$\rho_c^j = [\rho_c^{j-} , \rho_c^{j+}],$$

Thus, the resulted class prototype for the total number of features (m) is given by the vector of intervals

$$\boldsymbol{\rho_c=[\rho_c^1, \rho_c^2,, \rho_c^3,\ldots\ldots\ldots., \rho_c^m]}$$

k-means clustering Algorithm

K means is a popular numeric data clustering algorithm. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The basic steps of k-means algorithm are-

1. Choose k objects as the initial set of prototypes.
2. Partition the data using the prototypes.
3. Recalculate the cluster prototypes.
4. Repeat step 2 and 3 until there are no changes in the prototypes.

3. PROPOSED METHOD

The proposed method is based on the Relief algorithm. The Relief algorithm uses a dissimilarity margin based measure to put a weight for each feature to find whether that feature is important or not. Suppose a dataset D has n instances and m features then the Relief algorithm uses the following steps to feature selection-

Step1. Initiate the weights vector to zero: w[m] = 0.

Step2. Input an interval dataset D which has *n* numbers

of data with *m* numbers of features

Step3. 3.1. for j = 1…..m.

3.2. for each instance x present in D.

3.3. w_j=w_j+ ($dmiss_x^j$- $dhit_x^j$)

Step4. The chosen feature set is { j | $w_j > \tau$};

where τ is an input threshold.

For an instance x, $dmiss_x^j$ and $dhit_x^j$ (step 3.4) are calculated based on a dissimilarity margin as shown below-

Dissimilarity Margin1 (DisM1)

For a randomly selected instance x with m interval features, Relief searches for its two nearest neighbors: one from the same class, called *nearhit(x)*, and the other from the different class, called *nearmiss(x)*. Thus dissimilarity margin1 (DisM1) for the data x is defined as-

$$DisM1_x = \frac{1}{m}\sum_{j=1}^{m}(dmiss_x^j - dhit_x^j),$$

Where, $dmiss_x^j$= dist(x^j, nearmiss(x)^j)

$dhit_x^j$ = dist(x^j, nearhit(x)^j)

and *dist* denotes any distance measure available for the data.

To best of our knowledge till now, the Relief algorithm has not been used for feature selection in interval data. In this method we used this algorithm on interval data by considering *dist* as a distance measure for interval data. Again, the drawback of the Relief algorithm is that finding nearest neighbor is a computationally complex task if the neighborhood is large. So, in this proposed method we used two other dissimilarity margins in Relief algorithm, which uses class prototype instead of nearest neighbor and this will reduce the computational complexity of Relief algorithm. We assume that problem is a multiclass problem i.e more than one classes are present in the data. Thus, the proposed Dissimilarity margin is -

Dissimilarity Margin2 (DisM2)

For a data point x with m interval features Dissimilarity margin2 (DisM2) is defined as-

$$DisM2_x = \frac{1}{m}\sum_{j=1}^{m}(dmiss_x^j - dhit_x^j),$$

Where, $dmiss_x^j$= dist(x^j, $\rho_{near\_c_x'}{}^{j}$)

$dhit_x^j$ = dist(x^j, $\rho_{c_x}{}^{j}$)

Here, *dist* denotes any distance measure available for interval data, $\rho_{c_x}{}^{j}$ denotes jth component of the prototype vector of the class where x belongs, $\rho_{near\_c_x'}$ denotes the nearest prototype vector for x among the prototypes of the classes in which x does not belong and $\rho_{near\_c_x'}{}^{j}$ denotes jth components of $\rho_{near\_c_x'}$.

The nearest prototype vector $\rho_{near\_c_x'}$. is calculated in two ways-

DisM2_a

First the distances between x and all the prototypes of other classes (where x does not belong) are calculated and the prototype (suppose $\rho_{c_k'}$) to which distance is minimum is considered as $\rho_{near\_c_x'}$ i.e

$$D(x, \rho_{c_k'}) = minimum_{\forall c', x \notin c'} (D(x, \rho_{c'}))$$

Here, D(A, B) denotes distance between two interval patterns A and B and $\rho_{c'}$ denotes prototypes for classes where x does not belong.

DisM2_b

Secondly, the distances between jth components of x (x^j) and jth components of all the prototypes of other classes (where x does not belong) are calculated and the jth component of the prototype (suppose $\rho_{c_k'}$) to which the distance from x^j is minimum is considered as $\rho_{near\_c_x'}{}^{j}$. Thus, each jth components $\rho_{near\_c_x'}{}^{j}$ of $\rho_{near\_c_x'}$ is considered such that -

$$dist(x^j, \rho^j_{c_k'}) = minimum_{\forall c', x \notin c'} (dist(x^j, \rho^j_{c'}))$$

Here, dist(A^j, B^j) denotes distance between the jth components of two interval patterns A and B.

4. EXPERIMENTAL RESULTS

We develop programs in C++ to compare the performance of the proposed methods with the original Relief algorithm. We apply all these methods on three interval datasets, which are prepared from the real life data found in the meteorology site https://www.wunderground.com.

Dataset Description

To prepare the data we consider only 5 interval features namely 1. Temperature (min-max), 2. Dew point (min-max), 3. Humidity (min-max), 4. Sea level pressure (min-max), 5. Visibility (min-max) and an attribute 'Events' as the class attribute or decision attribute. Data containing missing values has not considered.

(a) **Guwahati Data:** In this dataset day wise weather information of Guwahati city has been included for the years 2005, 2006, 2007. It contains 476 data. Number of classes present is 12.

(b) **Delhi Data:** In this dataset day wise weather information of Delhi city has been included for the years 2005, 2006, 2007. It contains 444 data. Number of classes present is 6.

(c) **Guwahati-Delhi Data:** In this dataset day wise weather information of both Guwahati and Delhi city has been included for the years 2005, 2006, 2007. It contains 920 data. Number of classes present is 12.

Results analysis

We apply all the three methods i.e RDisM1(Relief using DisM1), RDisM2_a(Relief using DisM2_a), RDisM2_b(Relief using DisM2_b) on the above datasets using three different distance measures (L_1-norm, L_2-norm, L_∞-norm as given in section 3) to find the feature subsets. After that the k-means algorithm is used to cluster the data containing that features only. After that the clustering accuracy obtained from different methods are compared as shown in Table 1, Table 2, Table 3. The k-means apply for different values of k (from 2 to 10) and the minimum (Min), average (Avg) an maximum (Max) accuracies are given. It is well known fact that k-means is used for numeric data clustering purpose. So, here to cluster interval data using k-means, the prototype for each cluster is calculated using the method given in section 2 and data are assigned to its nearest prototype using the distance measure L_2-norm for interval data as given in section 3.

Table 1: Results obtained using distance measure L1-norm

Data	Method	Feature subsets	Accuracy using K-Means		
			Min	Avg	Max
Delhi	RDisM1	{1,2,3,4,5}	0.56	0.57	0.59
	RDisM2_a	{3}	0.59	0.6	0.62
	RDisM2_b	{1,2,3}	0.59	0.64	0.69

Guwahati	RDisM1	{1,2,3,4,5}	0.44	0.45	0.46
	RDisM2_a	{4}	0.44	0.45	0.48
	RDisM2_b	{1,2,3,4,5}	0.44	0.45	0.46
Guwahati Delhi	RDisM1	{1,2,3,4,5}	0.59	0.65	0.71
	RDisM2_a	{3,4}	0.53	0.54	0.56
	RDisM2_b	{1,2,3,4}	0.38	0.48	0.56

Table2: Results obtained using distance measure L2-norm

Data	**Method**	**Feature subsets**	**Accuracy using K Means**		
			Min	**Avg**	**Max**
Delhi	RDisM1	{1,2,3,4,5}	0.56	0.57	0.59
	RDisM2_a	{1,2,4,5}	0.59	0.66	0.74
	RDisM2_b	{4,5}	0.59	0.69	0.72
Guwahati	RDisM1	{1,2,3,4,5}	0.44	0.45	0.46
	RDisM2_a	{1,2,5}	0.44	0.44	0.45
	RDisM2_b	{5}	0.44	0.45	0.46
Guwahati Delhi	RDisM1	{1,2,3,4,5}	0.59	0.65	0.71
	RDisM2_a	{1,2,5}	0.56	0.56	0.57
	RDisM2_b	{5}	0.57	0.59	0.61

Table 3: Results obtained using distance measure L∞-norm

Data	**Method**	**Feature subsets**	**Accuracy using K means**		
			Min	**Avg**	**Max**
Delhi	RDisM1	{1,2,3,4,5}	0.56	0.57	0.59
	RDisM2_a	{3,4}	0.59	0.64	0.71
	RDisM2_b	{1,2,3}	0.59	0.64	0.69
Guwahati	RDisM1	{1,2,3,4,5}	0.44	0.45	0.45
	RDisM2_a	{3}	0.44	0.44	0.44
	RDisM2_b	{1,2,3,4}	0.44	0.45	0.46
Guwahati Delhi	RDisM1	{1,2,3,4,5}	0.59	0.65	0.71
	RDisM2_a	{3}	0.5	0.51	0.52
	RDisM2_b	{1,2,3,4}	0.38	0.48	0.56

In all the experiments the input threshold τ is set to 0 i.e a feature is selected if its weight is greater than equal to 0. Again all the features are numbered from 1 to 5 according to their order as mentioned above.

From Table 1, Table 2 and Table 3 it is seen that for all datasets proposed methods (RDisM2_a and RDisM2_b) select less number of features than RDisM1 using distance measures L_1-norm, L_2-norm, $L\alpha$-norm. On Delhi and Guwahati dataset, the feature subset selected by proposed methods give better minimum, average and maximum accuracies than RDisM1 for all the distance measures.

In Table1 it is seen that for Delhi dataset feature subset {1,2,3} gives maximum accuracies which is obtained by the proposed method RDisM2_b using L_1-norm. Again, on Delhi dataset using L_2-norm (Table 2) and $L\alpha$-norm (Table3) the proposed method RDisM2_a gives maximum accuracy with feature subset {4} and {3,4} respectively.

For Guwahati dataset feature subset {4} gives maximum accuracies which is obtained by the proposed method RDisM2_a using L1-norm (Table1). Again, on this dataset using L2-norm (Table 2), RDisM1 and the proposed method RDisM2_b gives maximum accuracy with feature subset {1,2,3,4,5} and {5} respectively, but the proposed RDisM2_b is better, because it selects less number of features. Using Lα-norm (Table3), The proposed method RDisM2_b selects features {1,2,3,4} for Guwahati dataset, which can obtain maximum accuracy.

For Guwahati-Delhi dataset RDisM1 selects features {1,2,3,4,5} for each distance measure and it gives maximum accuracy.

5. CONCLUSION

In this paper two feature selection methods based on Relief algorithm has been discussed. Relief is a feature selection algorithm mainly used for numeric data. Here, we used it for interval data by using some distance measures available for interval data. Also, unlike RDisM1 algorithm, the proposed methods (RDisM2_a and RDisM2_b) work on class prototype instead of nearest neighbor. From the analysis of results, it is seen that proposed methods selects less number of features for all datasets. The feature subsets which are selected by the proposed method for Guwahati and Delhi datasets are better than the feature subsets which are selected by RDisM1.

REFERENCES

[1] Guyon I., Elisseeff A., “An Introduction to Variable and Feature Selection”, *Journal of Machine Learning Research,* 3, 2003, pp 1157-1182

[2] Kilian Q. W., John B., Lawrence K. S., “Distance metric learning for large margin nearest neighbor classification”, *Advances in neural information processing systems,* 2005, pp 1473-1480

[3] Bachrach R. G., Navot A., Tishby N., “Margin based feature selection - theory and algorithms”, Proceedings of the twenty-first international conference on Machine learning (ICML '04), Alberta, Canada, July 04 - 08, 2004, pp 43-50

[4] Kira, K., Rendell, L., “A practical approach to feature selection”, Proc. 9th International Workshop on Machine Learning, 1992, pp. 249–256.

[5] Robnik-Sikonja M, Kononenko I, “Theoretical and empirical analysis of ReliefF and RReliefF”, Machine Learning 2003, 53, pp. 23-69.

[6] Hedjazi L., Martin J. A., Lann M. V. L, “Similarity-margin based feature selection for symbolic interval data”, Pattern Recognition Letters, 32, 2011, pp 578–585

[7] Roh J. W., Yi B. K., “Efficient indexing of interval time sequences”, Information Processing Letters, 109, 2008, pp 1–12