IJMI
International Journal of Machine Intelligence
ISSN: 0975–2927 & E-ISSN: 0975–9166, Volume 3, Issue 4, 2011, pp-349-353
Available online at http://www.bioinfo.in/contents.php?id=31

DIMENSIONALITY REDUCTION - A ROUGH SET APPROACH

SABU M.K.[1*], RAJU G.[2]
[1]Department of Computer Applications, M.E.S College, Marampally, Aluva-7, Kerala, India
[2]School of Information Science & Technology, Kannur University, Kannur, Kerala, India
*Corresponding Author: Email- sabu.mes@rediffmail.com

Received: November 06, 2011; Accepted: December 09, 2011

Abstract- In this paper we propose a novel approach of feature ranking for feature selection. This method is particularly useful for applications handling high dimensional datasets such as machine learning, pattern recognition and signal processing. This process is also applicable to small and medium sized datasets to identify significant features or attributes for a particular domain using the information contained in the dataset alone and hence the method preserves the meaning of the existing features. With the help of the proposed method, redundant attributes can be removed efficiently without sacrificing the classification performance. In this approach, after eliminating the outlier data elements from the dataset, features are ranked to identify the predominant features of the dataset. The discernibility matrix in RST is used as a tool to discover the data dependencies existing between various features and features are ranked based on these data dependencies. A method using Centre of Gravity (CoG) line is suggested to determine this discrimination frequency within a reduced computational effort. To evaluate the performance of the algorithm, we applied the proposed algorithm on a test dataset consisting of 3000 offline handwritten samples of 10 Tamil characters. The outcome of the experiment shows that the new method is efficient and effective for dimensionality reduction.
Key words – Rough Set Theory, Center of Gravity line, Discernibility Matrix, Degree of Discrimination.

Introduction

Rough Set Theory (RST), introduced by Z. Pawlak, is a mathematical tool to deal with uncertainty and vagueness in information systems using the granularity structure of the data [1]. If we have exactly the same information in two objects then we say that they are indiscernible (similar), which means we cannot distinguish them with known knowledge. These granules or group of similar objects are the basic building blocks for handling uncertainty [6]. The Rough Set approach provides efficient algorithms for finding out hidden patterns in data, minimal sets of data (data reduction), evaluating significance of data and generating sets of decision rules from data [1]. In RST all computations are performed directly on datasets. It requires no additional parameters to operate, such as a probability distribution in statistics or a grade of membership from fuzzy set theory etc., other than the supplied data. One advantage of RST is that it provides a well understood formal model which is very helpful in generating several kinds of information such as relevant features or association rules using minimal model assumptions [2]. The discernibility matrix in RST is useful for representing the knowledge regarding the discrimination between various objects of an information system. This degree of discrimination provides a new method of selecting important features automatically.

In this paper, we propose a novel method of attribute reduction, a problem encountered in many areas such as machine learning, pattern recognition and signal processing. When the number of features increases there is a chance of generating spurious patterns by the learning algorithm, that are not valid in general and this may reduce the performance of the learning algorithm. As a result, high dimensionality posed an open challenge for classification algorithms which lead to the design of many feature selection algorithms. The efficiency and accuracy of feature selection algorithms mainly depends on the quality of the input data. In the case of a dataset with outlier points, an algorithm for feature selection may produce irrelevant features. So, in attribute selection, the handling of these outlier points becomes an important issue. Hence, as a preprocessing step, a straight forward method is applied to remove the outlier points from the input dataset. To reduce the running time, in this approach, each class of objects in the given dataset is automatically replaced with a single representative vector. The vector representing each class is generated by employing the idea of a Centre of Gravity (CoG) line, an imaginary line upon which the CoG of a set of point lies [13]. Then attributes are ranked according to their relative significance in extracting knowledge by constructing a discernibility matrix of the reduced dataset. From the ranked attribute set, significant attributes are selected by specifying a pre-defined size for the number of attributes to be selected or by specifying a pre-defined minimum threshold value for the occurrence frequency. With this approach the complexity of the reduction process reduces from $O(n^2m)$ to $O(c^2m)$, where n is the total number of objects, c is the number of samples and m is the number of features in the dataset.

The rest of this paper is structured as follows. The procedure used to select the significant attributes is given in Section 2. Section 3 presents the proposed algorithm. In Section 4, we present a brief description of the dataset used for experimentation and the experimental results are also presented. Finally we conclude the paper in Section 5.

Ranking of Conditional attributes of a decision table

In the given information system all attributes are not equally important and by identifying predominant attributes and performing various data mining operations on these attributes will produce the same result as the original information system but with a reduced computational effort. Hence the complexity of the overall process can be greatly reduced. In RST this attribute selection process is done by using only the information available in the data set without requiring any additional information. This is based on the data dependencies existing in between various attributes in the data set [9, 10]. Rough set based attribute selection provides a method to reduce the amount of knowledge involved in the original data prior to any processing needed to retrieve the actual information. The purpose of feature selection is to identify the significant features, eliminate the irrelevant or dispensable features such that the resulting reduced data set is consistent with the decision attribute. This is helpful for building a good learning model by preserving the information content. The benefits of the feature selection are twofold: it considerably reduces the running time of the classification algorithms and increases the accuracy of classification [11].

Ranking of attributes according to their relative significance in extracting knowledge is an important issue in data analysis and decision making [3][5]. The process is also helpful for attribute reduction. The key idea of this attribute ranking process is borrowed from attribute reduction based on discernibility matrix in RST [8]. For this purpose the actual definition of discernibility matrix is slightly modified. This modification is mainly done by capturing the discrimination information involved in various object pairs. The advantage of this method is that this will work not only with dataset consisting of discrete attributes, but also with continuous attribute values. In order to handle continuous attribute values, the basic definition of discernibility matrix is modified using a distance function such as absolute distance [8].

In RST, the discernibility matrix is a symmetric $|U| \times |U|$ matrix, which can represent the discrimination information involved in all the conditional attributes of the given information system. Its entries C_{ij} can be defined as

$$C_{ij} = \begin{cases} a \in A \mid a(x_i) \neq a(x_j), \text{ if } d(x_i) \neq d(x_j) \\ \varnothing, \text{ otherwise} \end{cases} \quad (1)$$

To perform attribute ranking a modified discernibility matrix of size $m \times n$ is defined, where m is the number of object pairs (x, y) such that $d(x) \neq d(y)$ and n is the number of conditional attributes. The entries d_{ij} of the new matrix is defined as

$$d_{ij} = \begin{cases} 1, a_j(x) \neq a_j(y) \\ 0, \text{otherwise} \end{cases} \quad (2)$$

where (x, y) represents the i^{th} object pair O_i satisfying $d(x) \neq d(y)$ and j represents the index of the conditional attribute [12]. The column sum of this matrix gives the significance (frequency) of each conditional attribute. This significance value is proportional to the discrimination power of the attribute. Hence these frequency values play an important role in the ranking of conditional attributes leading to determine the significant attributes automatically [8].

The discernibility matrix obtained from a sample decision table shown in Table 1, where $\{u_1, u_2, u_3, u_4\}$ represents the given objects, $\{a_1, a_2, a_3\}$ represents the conditional attributes and d represents the decision attribute, is given in Table 2.

A '1' entry in the discernibility matrix shown in Table 2 indicates that the corresponding conditional attribute can discriminate the objects in the pair separately. After completing the matrix with the discernibility information, the significance (frequency) of each attribute can be computed by summing the corresponding column. The larger the sum is, more example pairs the attribute can discriminate, that is, the power of discrimination of that attribute is high. For example, according to Table 2, the significance of a_3 is 4; the significance of both a_1 and a_2 is 2. Hence a_3 is more significant compared to a_1 or a_2, for discriminating various objects of the decision table.

This discernibility matrix can very easily be generalized to handle continuous attribute values only if we adopt one kind of distance function, such as absolute distance [8]. Using this idea, the entries of the discernibility matrix are defined as

$$dij = \begin{cases} |a_j(x) - a_j(y)| & \text{if } a_j(x) \neq a_j(y) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where (x, y) represents the i^{th} object pair O_i satisfying $d(x) \neq d(y)$ and j represents the index of the conditional attribute.

With the help of this modified discernibility function, the degree of discrimination of various features are calculated as

$$Dist(i) = \sum_{j=1}^{U} dji \quad (4)$$

For the purpose of ranking various features, a discernibility matrix is constructed by applying formula (1). The significance of various features is then computed separately by summing the corresponding columns of the matrix and features are ranked based on this frequency value. This will provide a domain dependent approach to extract automatically, significant features representing the given knowledge base and eliminate unimportant ones

from the original high dimensional feature space with minimum information loss.

Proposed work

Consider a decision table T={U, A, d}, where U is a finite set of objects {x_1, x_2, …, x_n}, A is a finite set of conditional attributes and d is a finite set of decision attributes. Most of the real world data have apparent extraneous data points clearly not belonging to any of the classes and they are called outlier points. Detecting and removing these outlier points before performing any data mining operation is a good approach for generating valid patterns from the existing data. Here we consider a decision table with only one decision attribute. In this decision table, a filtering method is applied to improve the quality of the data. This filtering is done by removing the outlier points from the decision table by employing a straight forward method. One kind of solution to remove these outlier points is to select for processing only those objects x from each class of the decision table such that the attribute values of x lies in between [$\mu-\sigma$, $\mu+\sigma$], where μ and σ are the mean and variance of the considered attribute values in that class. To select these objects, for each class, we calculate the mean μ_i and standard deviation σ_i of each attribute (feature) a_i, where i=1, 2, 3… n and n is the total number of attributes in the dataset. Then for each object, select only those features a_i which satisfies $\mu_i - \sigma_i \le a_i \le \mu_i + \sigma_i$. If the total number of such selected features is greater than n/2, that object x is selected for processing; otherwise it is discarded. This process is repeated by recalculating the mean and variance of the class until we get a sufficiently stable set of objects for that class. The same strategy can be adopted for eliminating the outlier points of other available classes also.

In this method, ranking of attributes is mainly done by calculating the discrimination frequency of various attributes. Since the ranking is based on the discrimination frequency, a method is introduced here to determine the discrimination with in a reduced computational effort. For this purpose, in the original dataset, each class of objects is replaced with a vector representing the corresponding class. To determine the representative vector autonomously, the idea of a Center of Gravity (CoG) line is suggested. CoG line is an imaginary line in which the centre of gravity of a group of point lies [12]. For equally weighted distinct set of points, the CoG line will lie midway between them. When the distinction between various points in clusters varies ambiguously the line will move up or down to reflect the change. A reasonable vector representation of the set of points in each cluster is the line upon which their centre of gravity lies. The CoG line of a set of points in the plane is positioned such that the sum of all perpendicular distances from the points to this line is zero. These calculations may be weighted if the CoG line is seemingly distorted by outlier points [13]. Following this method, the vector representation v_i of each class of objects is obtained by selecting all those feature values m(x,c), x=1, 2, 3, …, n which minimizes the following sum of differences

$$\left|\sum_{j=1}^{n}(m(j,i)-w.m(x,i))\right|, \quad i=1,2,3,\ldots,p \qquad (5)$$

where n represents the number of objects in each cluster m, and p represents the total number of features [13].

For the purpose of ranking the features, a new decision table is constructed from the original table by replacing each class of objects with the corresponding CoG vector and by considering the labels of the training classes as the decision attribute value. Using this new decision table, discrimination of various attributes (features) is calculated by applying formula (1).

Based on the degree of discrimination it is possible to arrange the features of the original dataset in descending order. The degree of discrimination actually reflects the relative importance of various features in the knowledge domain. By setting a suitable threshold value for degree of discrimination or specifying the number of features needed for processing important features can be selected automatically from the large feature set.

Algorithm for Feature Selection

In this feature selection algorithm, the original decision table and the number of features to be selected are given as input and the algorithm outputs a feature subset with specified number of features.

Algorithm: Feature Selection

Input: The decision table A and the number n to select the first n significant features.

Output: n significant features.

Step 1: Input the original decision table.

Step 2: Sort the rows of the decision table in ascending order of the decision attribute values.

Step 3: Modify the original decision table by applying steps 4 to 10

Step 4: For each class of objects in the original decision table do

Step 5: Repeat steps 6 to 10 until a sufficiently stable class of objects is obtained.

Step 6: For each attribute a_i in the class compute the mean μ_i and standard deviation σ_i of the attribute values.

Step 7: For each object x in that class check whether the attribute values d_i for that object belongs to the interval $[\mu_i - \sigma_i, \mu_i + \sigma_i]$.

Step 8: Count the total number m of attribute values in the object x satisfying the above condition.

Step 9: If m<n/2, where n is the total number of attributes, delete the object x from the class.

Step 10: Go to step 4 to select the next class.

Step 11: From decision table obtained in step 5, generate a new reduced decision table M consisting of the representative vectors of each class of the decision table using expression 5.

Step 12: Construct an indiscernibility matrix D with entries

$$\begin{cases} |a_j(x)-a_j(y)| & \text{if } a_j(x)\neq a_j(y) \end{cases}$$

[dij] =

0 otherwise

where (x, y) represents the i^{th} object pair O_i satisfying d(x) ≠d(y) and j represents the index of the conditional attribute.

Step 13: Find the sum of each column of the matrix D, which gives the frequency of discrimination of each feature.

Step 14: Arrange the features according to the descending order of the discrimination frequency calculated in step 13.

Step 15: From the sorted list of features, select the required number of significant features.

With the help of the algorithm, the features available in the original dataset can be arranged in descending order and the order is determined by considering the relative importance of various features in the knowledge domain. By setting a suitable threshold value for discrimination frequency or by specifying the number of features needed for processing, significant features can be selected automatically from the large number of available features.

Experimental results

Experiments were performed using a dataset of 3000 handwritten character samples consisting of 10 Tamil characters with each character having 72 features. For extracting features of the character zero crossing method is used. In digital image processing a "zero crossing" is a point where the black to white transition of pixel in an image. For this experiment, we consider only isolated Tamil characters collected from different persons belonging to different age groups, qualification and profession. The collected documents are scanned at 300 DPI. Characters are segmented using projection histogram method, cropped and stored as bmp images. Each preprocessed character image is then divided into 36 equal blocks. For each block, black to white transition (zero crossing) of all pixels in each row and each column are separately calculated and then take the row sum and column sum of these values. For 36 blocks we get 36*2=72 values and these values are used as the features. Using the proposed algorithm these 72 features are ranked as per their degree of significance. To evaluate the efficiency of the method the classification accuracy of the original dataset with 72 features is determined first. Then from the set of ranked features, as a first stage, 12 lowest ranked features are eliminated and again the classification accuracy is determined. The process of eliminating the least significant features and evaluating the classification performance of the resulting dataset is continued until the dataset is reduced to 45 significant features. For classification two different classifiers from Weka Data Mining toolkit were used: Multi Layer Perceptron (MLP) and Radial Basis Function (RBF). Table 3 contains the classification accuracies given by the classifiers at different stages of attribute reduction. To determine the accuracy 10 fold cross validation is used.

In Table 3, the classification accuracy of the dataset with full 72 features and the accuracy of the dataset obtained after eliminating the least significant 12 features are same in MLP, but slightly higher in RBF. When we reduce the dataset once again by eliminating 5 least significant features, that is, when the number of features reduces to 55, the classifiers produce the same result as in the previous case. This shows that the proposed method of feature selection is effective and the least significant 17 features are redundant and can be removed without affecting the classification performance. From the results presented in Table 3, it is interesting to note that the classification accuracy given by a classifier not only dependent on the number of features but also on the significance of features. The results also show that there is a lot of redundancy in the dataset which may be removed using the proposed approach without any information loss.

Conclusion

In this paper, we have introduced a novel rough set based attribute selection process. The selection is mainly based on the importance of these features to discriminate various objects in the domain. For this purpose, after eliminating the outlier data points, the features are ranked based on the power of discrimination of these features. To determine the frequency of discrimination of the features within a reduced computational effort, a method using CoG line of a set of data points is suggested. From the experiments on the handwritten character dataset we conclude that in a dataset all the features are not equally important and the classification accuracy mainly depends on the quality of the extracted features. So selecting the predominant features from a dataset becomes necessary to perform the classification tasks optimally in the data mining process. Further investigation is needed to find out the representative vectors to replace each class of objects in the dataset so as to obtain the ranking of features more efficiently.

References

[1] Zdzislaw Pawlak (2002) *Information Sciences* 147 1 – 12, *Elsevier*.

[2] Richard Jensen and Qiang Shen (2009) *IEEE Transactions on Fuzzy Systems*, Vol. 17, No. 4.

[3] Jiye Li, Nick Cercone (2006) *KDM Workshop, Waterloo, Canada* .

[4] Jiawei Han and Micheline Kamber: *Elsevier*.

[5] Jiye Li(2007) *Ph.D thesis* from Internet.

[6] Zdzislaw Pawlak: *Kluwer Academic Publishers*.

[7] Oded Maimon and Lior Rokach: *Springer*.

[8] Songbo Tan, Yuefen Wang and Xueqi Cheng(2008) *ACM*.

[9] Qiang Shen, Alexios Chouchoulas (2001) *International Journal of Applied Mathematics and Computer Sciences*, Vol.11, No.3, 583-601.

[10] Richard Jensen(2005) *Ph.D thesis* from Internet.

[11] Thangavel K., Peethalakshmi A. (2009) *A review, Applied Soft Computing, 9,* 1-12.

[12] Ramadevi Yellasiri, C.R. Rao, Vivekchan Reddy: *Journal of Theoretical and Applied Information Technology*.

[13] Charlotte Bean, Chandra Kambhampati: (2008) *International Journal of Automation and Computing*, 05(1), 90-102.

Table 1: A sample decision table:

U	a_1	a_2	a_3	d
u_1	True	True	Very_high	1
u_2	False	True	Normal	0
u_3	False	False	High	0
u_4	False	True	Very_high	1

Table 2: The modified discernibility matrix of Table1:

Object pairs	a_1	a_2	a_3
(1, 2)	1	0	1
(1, 3)	1	1	1
(2, 4)	0	0	1
(3, 4)	0	1	1

Table 3: Classification accuracy given by the classifiers:

No. of features	Accuracy	
	MLP	RBF
72	97.9667	93.7
60	97.9667	93.8333
55	97.9667	93.7667
45	97.1	92.7667