Feature Selection based on Information Gain

B.Azhagusundari, Antony Selvadoss Thanamani

Abstract- The attribute reduction is one of the key processes for knowledge acquisition. Some data set is multidimensional and larger in size. If that data set is used for classification it may end with wrong results and it may also occupy more resources especially in terms of time. Most of the features present are redundant and inconsistent and affect the classification. In order to improve the efficiency of classification these redundancy and inconsistency features must be eliminated. This paper discusses an algorithm based on discernibility matrix and Information gain to reduce attributes.

Keywords: Attribute Reduction, Discernibility matrix, Information Gain

I. INTRODUCTION

Data collection and storage capabilities during the past decades have led to an information overload in all the fields especially in science. Researchers working in domains as diverse as engineering, medical, astronomy, remote sensing, economics, and consumer transactions face larger and larger observations and simulations on a daily basis. Such datasets that have been studied extensively in the past present new challenges in data analysis. Traditional statistical methods break down partly because of the increase in the number of observations which in turn lead to change in dimensions.

The dimension is the number of variables that are measured on each observation. High-dimensional datasets present many mathematical challenges as well as some opportunities and are bound to give rise to new theoretical developments.

One of the problems with high-dimensional datasets is that, in many cases, not all the measured variables are important for understanding the underlying phenomena of interest. While certain computationally expensive novel methods can construct predictive models with high accuracy from high-dimensional data. It is still of interest in many applications to reduce the dimension of the original data prior to any modelling of the data. Feature selection is the method that can reduce both the data and the computational complexity. Dataset can also get more efficient and can be useful to find out feature subsets.

II. RELATED WORK

Rough set based reduction [1] was proposed by Wa'el M. Mahmud, Hamdy N.Agiza, and Elsayed Radwan. The main contribution of this paper was to create a new hybrid model RSC-PGA (Rough Set Classification Parallel Genetic Algorithm) presented to address the problem of identifying important features in building an intrusion detection system. Tests has been carried out using KDD-99 dataset.

Manuscript received on January, 2013.

B.Azhagusundari, Ph.D Research Scholar, Nallamuthu Gounder Mahalingam College, Pollachi, Coimbatore, India.

Dr.Antony Selvadoss Thanamani, Professor and Head, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, Coimbatore, India

Rough set and neural network based reduction was proposed by Thangavel .K, & Pethalakshmi .A [2].This paper describes the reduction attribute with the help of medical datasets.

Protocol based classifications was proposed by", Kun-Ming Yu, Ming-Feng Wu, and Wai-Tak Wong [3]. This paper described the protocol based classification by using genetic algorithm with logistic Regression and implemented by KDD 99 dataset.

Data Analysis methodologies were described by,Shaik Akbar, Dr.K.Nageswara Rao Dr.J.A.Chandulal [4]. This paper describes so many methodologies such as rule based, machine learning and AI.

Discernibility matrix was described by Chuzhou [5] . This paper has a neat explanation about the discernibility matrix function and reduction of features.

Misuse and Anomaly detection using SVM, NBayes, ANN and ensemble approaches are discussed by T.Subbulakshmi[6]. This paper notifies the detection rate and false alarm rates. Multilayer Perceptrons, Naïve Bayes classifiers and Support vector machines with three kernel functions are used for detecting intruders. The Precision, Recall and F- Measure for all the techniques are calculated.

A rough set theory is a new mathematical tool to deal with uncertainty and vagueness of decision system and it has been applied successfully in all the fields by Y.Y.Yao and Y. Zhao, [7] . It is used to identify the reduct set of the set of all attributes of the decision system. The reduct set is used as pre-processing technique for classification of the decision system in order to bring out the potential patterns or association rules or knowledge through data mining techniques.

III. METHODOLOGY

3.1 Pre-processing

Attributes in the any datasets had all forms continuous, discrete, and symbolic, with significantly varying resolution and ranges. Most pattern classification methods are not able to process data in such a format. Hence pre-processing was required before pattern classification models could be built. Pre-processing consisted of two steps: The first step involved is mapping symbolic-valued attributes to numeric-valued attributes and the implemented non-zero numerical features.

S no	Age	Income	Student	Credit	Class
1	youth	high	no	fair	No
2	youth	high	no	excellent	No
3	middle	high	no	fair	Yes
4	senior	medium	no	fair	Yes
5	senior	low	yes	fair	Yes
6	senior	low	yes	excellent	No

7	middle	low	yes	excellent	Yes
8	youth	medium	no	fair	No
9	youth	low	yes	fair	Yes
10	senior	medium	yes	fair	Yes
11	youth	medium	yes	excellent	Yes
12	middle	medium	no	excellent	Yes
13	middle	high	yes	fair	Yes
14	senior	medium	no	excellent	No

Table 1: Example Dataset

In pre processing step the example dataset (Table 1) is given the attribute value for Age youth=0,middle=1 and senior=3 likewise the value for income high=0,medium=1 and low=2 , the value for student no=0 and yes=1, the value for credit fair=0 and excellent=1, the value for class No=0 and Yes=1.After pre processing the example dataset Table 1 becomes Table 2.

Sno	Age	Income	Student	Credit	Class
1	0	0	0	0	0
2	0	0	0	1	0
3	1	0	0	0	1
4	2	1	0	0	1
5	2	2	1	0	1
6	2	2	1	1	0
7	1	2	1	1	1
8	0	1	0	0	0
9	0	2	1	0	1
10	2	1	1	0	1
11	0	1	1	1	1
12	1	1	0	1	1
13	1	0	1	0	1
14	2	1	0	1	0

Table 2 : Pre-Processed Data

3.2 Feature Selection

Data sets for analysis may contain hundreds of attributes, many of which may be irrelevant to the mining task, or redundant. Although it may be possible for a domain expert to pick out some of the useful attributes, this can be a difficult and time consuming task, especially when the behaviour of the data is not well known. Leaving out relevant attributes or keeping irrelevant attributes may be detrimental, causing confusion for mining algorithm employed.

Thus the dimensionality reduction reduces the data size by removing such attributes from it. The method called attribute subset selection is applied to reduce the data size.. The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Mining on a reduced set of attributes has an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.

3.3 Discernibility matrix

Let (U,A) be an information system , M be a set of {T} called the discernibility matrix of (U,A) , such that T={ a € A, where C[I]≠C[J] } I,J=1,2,….n . The physical meaning of the matrix element M(x, y) is that objects x and y can be distinguished by any attribute in M(x, y). The pair (x, y) can be discerned if M(x, y) ≠ 0. A discernibility matrix M is symmetric, i.e., M(x, y) = M(y, x), and M(x, x) = 0. Therefore, it is sufficient to consider only the lower triangle or the upper triangle of the matrix.

Discernibility Function: For an information system (U,A), s discernibility function F is a boolean function of m Boolean variables $a_1,a_2……..a_m$ corresponding to the attributes $a_1,a_2,….a_m$ respectively, and defined as follows:

$F(a_1,a_2….a_m)=T_1 \square T_2 \square …..T_n$ where a_i € T

Absorptive law: Let G and H be two logical formulas, G □ (G □ H)=G □ (G □ H)=G. use distributive law(Multiplication) and absorptive law to transform the discernibility function into a disjunctive normal form(DNF), which is just the reduct space of the given information system. Every clause in this DNF is a reduct.

3.4 Proposed algorithm based Attribute Reduction

Step 1: Compute discernibility matrix for the selected dataset.

By using M[I,J]={ a € A, where C[I]≠C[J] and D[i]≠D[J]} I,J=1,2,….n ·····▶ Eq1

Where C are conditional attributes and D is a decision attribute. This discernibility matrix M is symmetric. Where M[x,y]=M[y,x] and M[x,x]=0. Therefore, it is sufficient to consider only the lower triangle or the upper triangle of the matrix.

Step 2: Compute the discernibiliy function for the discernibility matrix M[x,y] by using

F(x) = □ {□ M[x,y] / x,y € U; M[x,y] ≠0}..▶ Eq2

Step 3: Select the attribute, which belongs to the large number of conjunctive sets, numbering at least two, and apply the expansion law.

Step 4: Repeat steps 1 to 3 until the expansion law cannot be applied for each component.

Step 5: Substitute all strongly equivalent classes for their corresponding attributes.

Step 6: Calculate the Information gain for the simplified discernibility function contained attributes by using

The information gain is defined as

$Gain(S_j) = E(P_j) - E(S_j)$ ·····▶ Eq.3

Where

$$E(P) = \sum_{i=1}^{n} Pi \log_2 Pi$$

·····▶ Eq. 4

$$\sum_{i=1}^{n} Pi \log_2 Pi = -\frac{p1}{p} \log_2 \frac{p1}{p} - \frac{p2}{p} \log_2 \frac{p2}{p} \ldots \frac{pn}{p} \log_2 \frac{pn}{p}$$

·····▶ Eq.5

Where P_i is the ratio of conditional attribute P in dataset. When S_j has $|S_j|$ kinds of attribute values and condition attribute P_i partitions set P using attribute S_j, the value of information $E(S_j)$ is defined as

$$E(S_j) = \sum_{i=1}^{sj} I_j * E(Y_j)$$

·····▶ Eq.6

International Journal of Innovative Technology and Exploring Engineering
IJITEE
www.ijitee.org
Exploring Innovation

Step 7: Choose the highest Gain value and add it to the reduction set, and remove the attribute from the discernibility function.

Goto step 6 until the discernibility function reaches null set.

IV. RESULTS AND DISCUSSION

Effective input attributes selection from intrusion detection datasets is one of the important research challenges for constructing high performance IDS. Irrelevant and redundant attributes of intrusion detection dataset may lead to complex intrusion detection model as well as reduce detection accuracy.

Calculate the discernability matrix for the example dataset shown in (Table 3)

The discernibility function (DF) is defined as

DF=A □ D □ (A V B) □ (B V C) □ (A V D) □ (A V C) □ (C V D) □ (A V B V C V D) □ (A V B V C) □ (B V C V D) □ (A V B V D) □ (A V C VD)

Here A-Age, B-Income, C-student,D-Credit.

To calculate the information entropy by using the equation 3,4,5 and 6

Gain (Age)=0.2467

Gain (Income)=0.0292

Gain(Student) =0.1518

Gain(Credit)=0.0481

Choose the highest Gain value for the above listed values ie Gain(Age)=0.2467 and add it to the reduction set. And remove the attribute from the discernibility function DF . R={Age}.

DF=D □ (B V C) □ (C V D) □ (B V C V D)

Choose the next highest Gain value for the above listed values ie Gain(student)=0.1518 and add it to the reduction set. And remove the attribute from the discernibility function DF ie R={Age,Student}.

DF=D

The discernibility function contain only D so add it to the reduction set ie R={Age,Student ,Credit}.the discernibility function is empty ie DF=null. Stop the reduction.

Confusion matrix	**Predicted label**	
Actual Label	True Negative(TN)	False positive(FP)
	False Negative(FN)	True Positive(TP)

Table 3: Confusion matrix

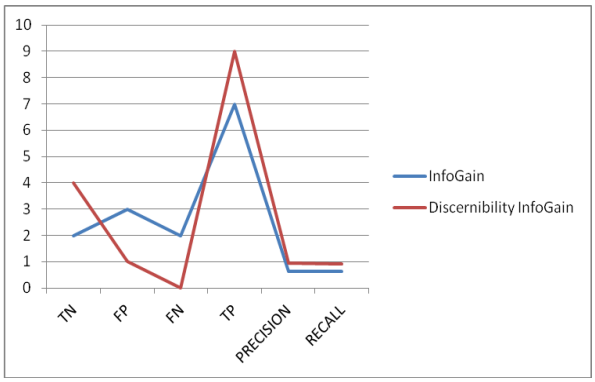

Fig: 1 :InfoGain Vs discernibility InfoGain

In this Fig: 1 the x-axis parameters are True Negative ,False Positive,False Negative, True Positive, precision and Recall, y-axis parameters are Information Gain and discernibility matrix with Information Gain. The curve indicates the comparsion result for these two methods.

V. CONCLUSION

In this work, this approach for selecting the best discriminate features using discernibility matrix and information gain is presented. From the Results of Table 4 the classification with selected features shows better results. The selection method using Information Gain and Discernibility shows better results in terms of number of features selected and accuracy than applying methods individually.

In future, this approach can be tested for other type data sets and to explore the possibilities of other methods of selecting optimal feature set. This would lead to the identification of the different methods of classification which will improve the results. The algorithm can be applied to different type of data sets which are required for dimensionality reduction and classification.

Table 3 : Discernibility Matrix for the example dataset

Method of selection	Features selected	Actual		Predicted				Accuracy	Recall
		Yes	No	TN	FP	FN	TP		
	All	9	5	2	3	2	7	0.629	0.643
Information Gain	All	9	5	2	3	2	7	0.629	0.643
Discernibility matrix and Information Gain	Age,student,credit	9	5	4	1	0	9	0.936	0.929

Table 4: Accuracy and recall calculation

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1		-	A	A,B	A,B,C		A,B,C, D		B,C	A,B,C	B,C, D	A,B, D	A,C	
2			A,D	A,B, D	A,B,C, D		A,B,C		B,C,D	A,B,C, D	B,C	A,B	A,C,D	
3	A	A,D				A,B,C, D		A,B						A,B,D
4	A,B	A,B,D				B,C,D		A						D
5	A,B, C	A,B,C, D				D		A,B,C						B,C,D
6			A,B,C, D	B,C, D	D		A		A,D	B,D	A,B	A,B, C	A,B,D	
7	A,B, C,D	A,B,C				A		A,B,C, D						A,B,C
8			A,B	A	A,B,C		A,B,C, D		B,C	A,B,C	C,D	A,D	A,B,C	
9	B,C	B,C,D				A,D		B,C						A,B,C, D
10	A,B, C	A,B,C, D				B,D		A,B,C						C,D
11	B,C, D	B,C				A,B		C,D						A,C
12	A,B, D	A,B				A,B,C		A,D						A,C
13	A,C	A,C,D				A,B,D		A,B,C						A,B,C, D
14			A,B,D	D	B,C,D		A,B,C		A,B,C, D	C,D	A,C	A,C	A,B,C, D	

REFERENCES

[1] Wa'el M. Mahmud, Hamdy N.Agiza, and Elsayed Radwan (October 2009) ,Intrusion Detection Using Rough Sets based Parallel Genetic Algorithm Hybrid Model, Proceedings of the World Congress on Engineering and Computer Science 2009 Vol II WCECS 2009, San Francisco, USA

[2] Thangavel, K., & Pethalakshmi, A. Elseviewer (2009)., Dimensionality reduction based on rough set theory *9*, 1-12. doi: 10.1016/j.asoc.2008.05.006.

[3] Kun-Ming Yu, Ming-Feng Wu, and Wai-Tak Wong (April,2008), Protocol-Based Classification for Intrusion Detection, APPLIED COMPUTER & APPLIED COMPUTATIONAL SCIENCE (ACACOS '08), Hangzhou, China.

[4] Shaik Akbar, Dr.K.Nageswara Rao , Dr.J.A.Chandulal (August 2010),Intrusion Detection System Methodologies Based on Data Analysis, International Journal of Computer Applications (0975 – 8887) Volume 5– No.2.

[5] Chuzhou University, China, Guangshun Yao,Chuanjian Yang,1Lisheng Ma, Qian Ren (June 2011) An New Algorithm of Modifying Hu's Discernibility Matrix and its Attribute Reduction, International Journal of Advancements in Computing Technology Volume 3, Number 5.

[6] T. Subbulakshmi,A. Ramamoorthi, and Dr. S. Mercy Shalinie9August 2009), Ensemble design for intrusion detection systems, International Journal of Computer science & Information Technology (IJCSIT), Vol 1, No 1.

[7] Y.Y.Yao and Y. Zhao (2009), Discernibility matrix simplication for constructing attribute reducts, Information Sciences, Vol. 179, No. 5, 867-882.

B.Azhagusundari received her B.Sc Mathematics and Master of Computer Applications from NGM College, Pollachi, Coimbatore, India. She completed her Master of Philosophy in Bharathidasan University, Trichy. Presently she is working as an Assistant Professor in the P.G Department of Computer Applications in NGM College (Autonomous), Pollachi. Her area of interest includes data Mining and Networking. Now she is pursuing her Ph.D Computer Science in Mother Teresa University, Kodaikannal.

Dr. Antony Selvadoss Thanamani is presently working as Professor and Head, Dept of Computer Science, NGM College, Coimbatore, India (affiliated to Bharathiar University, Coimbatore). He has published more than 100 papers in international/ national journals and conferences. He has authored many books on recent trends in Information Technology. His areas of interest include E-Learning, Knowledge Management, Data Mining, Networking, Parallel and Distributed Computing. He has to his credit 24 years of teaching and research experience. He is a senior member of International Association of Computer Science and Information Technology, Singapore and Active member of Computer Science Society of India, Computer Science Teachers Association, New York