IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

A Comparative Study of Feature Selection Algorithms in Data Mining

P.Kavipriya[1], Dr.K.Karthikeyan[2]

Asst Professor, Dept of computer Science & Application & SS, Sri Krishna Arts and Science College, Coimbatore[1]

Head, Department of Computer Science, Government Arts College, Palladam[2]

Abstract: Data mining is the method of extraction of related data from a collection of large dataset. Mining of a fastidious data related to a concept is done on the basis of the feature of the data. The accessing of these features thus for data retrieval can be termed as the feature selection mechanism. Different types of feature selection methods are being used. Feature selection methods in data Mining problem aim at selecting a subset of the features, which illustrate the data in order to acquire a more necessary and compact representation of the available information. The preferred subset has to be small in size and must maintain the information that is most useful for the specific application. This paper try to analyze Feature Selection algorithms clearly with the purpose to examine strengths and weaknesses of some widely used Feature Selection methods.

Key words: Feature selection algorithm, Euclidian distance, T-test, Information gain, Markov blanket filter.

I. INTRODUCTION

With the rapid development of data acquisition and network technology, data mining is widely used in all areas of society [1]. It predicts future trends, behaviors and knowledge-driven decision. Data mining is a process of knowledge discovery. The KDD is an automated process of knowledge discovery from the original data. The KDD include some essential steps as follows,

- Data cleaning,
- Data integration,
- Data selection,
- Data transformation,
- Pattern evaluation
- Knowledge representation.

Among the steps the data selection is very much important to select the relevant feature and remove the irrelevant attributes. Feature selection is one of the data mining techniques used to discover the unknown class [1]. Fig 1. Shows the steps in data mining process.

The feature selection algorithm eliminates the unrelated and repeated features from the original dataset to develop the classification accuracy. The feature selections also reduce the dimensionality of the dataset. It increases the learning accuracy, improving result comprehensibility. The feature selection avoid over fitting of data. The feature selection also known as attributes selection which is used for best partitioning the data into individual class.

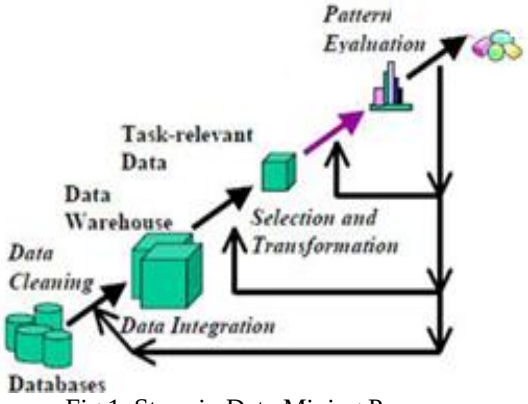

Fig 1. Steps in Data Mining Process

DOI 10.17148/IJARCCE.2017.61109

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

II.FEATURE SELECTION

Feature as a set for suitability is estimated by a subset selection a subset of features. Feature subset selection methods are categorized into following methods as,

- Wrappers
- Filters
- Embedded
- Hybrid methods

It has been a dynamic and productive field of research area in pattern recognition, machine learning, statistics and data mining communities [2]. The major aim of feature selection is to select a subset of input variables by reducing features, which are unrelated and redundant information. It has confirmed in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned outcomes [3].

Feature selection in supervised learning has a major objective of finding a feature subset that generates higher classification accuracy. As the dimensionality of a domain expands, the number of features N increases. Finding a best feature subset is difficult and problems related feature selections have been proved to be NP-hard [4]. It is essential to describe traditional feature selection process, which consists of four basic steps [5], namely,

- Subset generation,
- Subset evaluation,
- Stopping criterion,
- Validation

Subset generation is a search system that produces candidate feature subsets for evaluation based on a definite search strategy. All candidate subset is estimated and measure up to with the previous best one according to a certain evaluation. If the fresh subset comes to be better, it replace best one. This procedure is repetitive until a given stopping state is satisfied. Grading of features decides the significance of any individual feature, ignoring their potential interactions. Grading methods are based on statistics, information theory, or on some functions of classifier's outputs [6].

Algorithms for feature selection divided into two wide classes namely wrappers that apply the learning algorithm itself to estimate the values of features and filters that assess features according to heuristics based on common characteristics of the data. A number of justifications for the use of filters for subset selection have been discussed [7] and it has been reported that filters are comparatively faster than wrappers. Like Decision Tree, Bayesian Network, and other classification algorithms have also been discussed [8]. But, they expose only classifier accuracy without performing the feature selection procedures.

There are four basic operations in the feature selection method (Fig 1):

i. A generation procedure to create the next candidate subset,
ii. An evaluation function to estimate the subset under examination,
iii. A stopping criterion to choose when to stop and
iv. A validation procedure to verify whether the subset is suitable [9].

Fig.2 Feature Selection Process

DOI 10.17148/IJARCCE.2017.61109

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

III. BASIC FEATURE SELECTION ALGORITHM

Input:
S - Data sample f with features X, |X| = n
J - Evaluation measure to be maximized
GS – successor generation operator
Output:
Solution – (weighted) feature subset
L: = Start Point(X);
Solution: = {best of L according to J };
Repeat
L: = Search Strategy (L, GS (J), X);
X':= {best of L according to J};
If J (X') =J (Solution) or (J (X') =J (Solution) and |X'| < |Solution|) then
Solution: =X';
Until Stop (J, L).

The filter technique utilizes the discriminating criterion for feature selection. The correlation coefficient and statistical test is used to filter the features in the filter feature selection technique [9]. The FSDD, RFS, CFS are the feature selection algorithm which utilizes the filter methodology. The relevance score is calculated for the features to ensure the correlation between the features. The calculated score is high with some threshold value then the particular feature is selected for further classification. When the ranking is small those feature are removed. This technique is very easy, quick and autonomous of classification algorithm. The followings are the basic filter feature selection algorithms,

S.NO	Feature Selection Algorithm
1	χ^2 test
2	Euclidian distance
3	T-test
4	Information gain
5	CFS-correlation based feature selection method
6	MBF- Markov blanket filter
7	FCBF-fast correlation based feature selection

1. χ2 Test

The chi-squared test is feature selection methodology used in filter technique. The chi squared statistical test ensures the independence between the two events. If X, Y are two events then the statistical independence is denoted by the following equations

$$P(XY) = P(X)\,P(Y) \text{ or}$$
$$P(X/Y) = P(X) \text{ and } P(Y/X) = P(Y)$$

The null hypothesis intimates that there is no correlation among the events. The events in the classification indicate the class [9].

2. Euclidian Distance

The Euclidian distance is a feature selection method used in filter method. In this process, the correlation between the features is calculated in terms of Euclidian distances. If there are n number of features in a sample feature says 'a' is measure up to with other n-1 features by calculating the distance among them using the following equation.

$$d(a,b) = \{\Sigma i\,(ai - bi)2\ \}½$$

The addition of new feature will not involve the distance among any two samples [9].

3. T-Test

The filter system uses the t-test for calculating the relationship among the two samples by comparing its mean value. The t-test utilizes the following formula to evaluate the mean value.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

$$T = \frac{X' - Y'}{\sqrt{\frac{S_x^2}{n1} + \frac{S_y^2}{n2}}}$$

The answer of the formula is ratio which points out the difference among the two mean values [10].

4. Information Gain

The entropy and the information achieve is an attribute measure which points out how much percentage the given attribute detach the training dataset according to their last classification. The Entropy for a set S is evaluate as

$$Entropy(s) = \sum_{i=1}^{n} -Pilog2Pi$$

Where 'n' is the number of classes, and the Pi is the probability of S belongs to class i. The achieve of A and S is calculated as

$$\text{Gain(A)=Entropy(S)} - \sum_{k=1}^{m} \frac{|Sk|}{|S|} \text{ x Entropy}(s_k)$$

S_k is the subset of $_S$ [11].

5. CFS

CFS is a Correlation-based Feature Selection algorithm which utilizes the filter technique for choosing the attributes. It is illustrated by Hall Correlation. The CFS algorithm applies a heuristic which measures the usefulness of individual features for predicting the class label along with the level of inter-correlation among them. The extremely correlated and unrelated features are avoided. The equation used to filter out the unrelated, unnecessary feature which leads the low down prediction of the class is calculate using the equation

$$Fs = \frac{N\overline{r_{ci}}}{N+N(N-1)\overline{r_{ii}}}$$

- N is the number of features in the subset,
- r_{ci} is the mean feature correlation with the class and
- r_{ii} is the average feature inter-correlation.

For calculating the correlations necessary for equation a number of information based measures of association were projected such as: the uncertainty coefficient, the gain ratio or the minimum description length principle. The best results achieved with the gain ratio used for feature-class correlations and symmetrical uncertainty coefficients used for feature inter correlations [11].

6. MBF

The Markov Blanket of a feature is applied to remove the irrelevant features from the feature set.

- Let G is the subset of the feature set S.
- Fi is the feature in G.
- M is some other subset independent of Fi.
- M is a Markov Blanket of Fi, if Fi is conditionally independent of G-M-Fi. If M is the Markov Blanket of Fi then Fi can be removing from the set G. The remaining features are denoted by

$$G'=G-Fi$$

The algorithm steps for MBF filter method [1]

Step1. Initialize G = F

Step2. Iterate

Step3. For each feature Fi $\in$ G,let Mi be the set of k feature Fj $\in$ G $\in$ {Fi} for which the correlations between Fi and Fj are the highest.

Step4. Compute Δ (Fi |M) for each i

Step5. Choose the i that minimizes Δ Fi |M) and define G=G -{Fi}

The Markov blanket filtering decreases the discrepancy among the conditional distributions by using conditional entropy [12].

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

7. FCBF

The FCBF algorithm for selecting the features using the filter method is as follows [12]

Step1. S is the set of candidate predictors,
M = Ø is the set of selected
Predictors
Step2. Searching X* (among S) which
Maximizes its correlation with Y→ρ
Step3. If ρ y,x*>= δ add X* into M and
Remove X* from S
Step4. Remove also from S all the variables
X such ρx,x*>= ρ y,x*
Step5. If S ≠Ø then GOTO (2), else
Step6. Stop

The algorithm sustains a dataset of extremely huge number of candidate predictors.

There are different kinds of feature selection algorithms. Some of them based on filter technique and some based on wrapper technique and some are based on embedded technique. Not all the feature selection sustains multiclass dataset. Some technique support only binary dataset. When the feature selection is used on high dimensional medical dataset the algorithm which selects the suitable and best features is not remember to increase the accuracy limitation of the classifier. The feature selection algorithm which sustains both binary dataset and the multiclass dataset sometimes generates high accuracy on the binary dataset but gives low accuracy when it is used in the multiclass data set. The feature selection algorithm must sustain multiclass dataset and generate high accuracy when applied on classification [10].

IV. ADVANTAGES AND DISADVANTAGES OF FEATURE SELECTION ALGORITHM

The advantages and disadvantages of feature selection algorithm in data mining are given in Table I.

Algorithms	*Advantages*	*Disadvantages*
χ2 test	Better model understandability and visualization Generalization of the model and reduced over fitting, as a result better learning accuracy is achieved. Efficiency in terms of time and space complexity for both training and execution time.	Ignoring the specific heuristics and biases of the classifier might lower the classification accuracy.
Euclidian distance	Time complexity is O(n), which is low as compared to other methods.	High sensitivity to noise and outliers, demand for extensive data preprocessing if to be applied as time series similarity measure. In spite of its merits, the Euclidean distance is not similarity measure of choice for time series
T-test	Eliminate subject-to-subject variability Control for extraneous variables No need large sample dataset	Ignores feature dependencies. Ignores interaction with the classifier.

Information gain	Eliminates redundancy Tests the relevance of features in combination with other features	Time complexity more as compared to filter methods O(nm2).
CFS-correlation based feature selection method	Tests the predictive power of genes Less computational complexity compared to other method Less prone to overfitting	Heavily dependent on the model, so they can fail to fit the data well.
MBF- Markov blanket filter	It easily scale to very high-dimensional datasets Computationally simple and fast, and they are independent of the classification algorithm. Feature selection wants to be performed only once, and then different classifiers can be estimated.	It ignores the communication with the classifier. Ignoring feature dependencies, which may lead to poor classification performance when compared to other types of feature selection techniques.
FCBF-fast correlation based feature selection	A feature goodness measure for classification. First, it helps remove features with near zero linear correlation to the class. It helps to reduce redundancy among selected features.	Slower than univariate techniques. Less sclable than univariate techniques, Ignores interaction with the classifier.

V. LITERATURE REVIEW

The author [13] introduces an algorithm for filtering information based on the Pearson $\chi 2$ test approach has been implemented and tested on feature selection. This is useful for high dimensional data where no sample set is large. This test is frequently used in biomedical data analysis and used only for nominal (discretized) features. This algorithm has only one parameter, statistical confidence level that two distributions are identical. Empirical comparisons with four other features selection algorithms (FCBF, CorrSF, ReliefF and ConnSF) are done to find quality of feature selected. This algorithm work fine with the linear SVM classifier.

Veerabhadrappa and Lalitha Rangarajan [14] designed a hybrid method to extract the features. In this method they used multi-level process to extract the important features. In the first level they used statistical method to extract the best features and in the second level they analysed the quality of the individual features which are extracted in the first level. Finally, based on the features quality measure the best features are extracted.

Sandya et al. [15] developed a new feature extraction method using fuzzy logic. In this method the fuzzy system generates a fuzzy score. This score is used to extract the most relevant features. They found that this method extract the efficient features and shows the better classification accuracy.

Tomasz Kajdanowicz et al. [16] developed a new method for feature extraction. In this method the new features are calculated by combining the network structure information and the class label. This method is able to extract the important features and show small improvement in the classification accuracy.

Gladis Pushpa Rathi and Palani [17] used the most common feature extraction technique to extract the features. In their research they used PCA and LDA to extract the most relevant features. This set of newly obtained features is applied to a Support Vector Machine (SVM) classifier and it shows improved classification accuracy.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

VI. CONCLUSION

The feature selection algorithms must select the applicable features and also remove the unrelated and conflicting features which cause the degradation of accuracy of the classification algorithms. The classification and feature selection algorithms must carry both binary as well as multiclass datasets. This paper describes the some basic feature selection algorithms in data mining. Although, it is not promising to declare that one approach is universally better compared to other methods. Their advantages and disadvantages were also discussed. Every method has its own advantages and disadvantages and performs differently on different datasets. The various algorithms are compared based on their common performance.

REFERENCES

[1] Ming-Syan Chen, Jiawei Han, Philip S yu. Data Mining: An Overview from a Database Perspective[J]. IEEE Transactions on Knowledge and Data Engineering, l996, 8(6):866-883.This paper shows the different feature selection algorithms.
[2] P. Mitra, C. A. Murthy and S. K. Pal. "Unsupervised feature selection using feature similarity," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301–312, 2002.
[3] H. Almuallim and T. G. Dietterich. "Learning boolean concepts in the presence of many irrelevant features," Artificial Intelligence, vol. 69, no. 1-2, pp. 279–305, 1994.
[4] K. R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, nos.1-2, pp. 273-324, 1997.
[5] M.Dash and H.Liu, "Feature Selection for Classification," An International Journal of Intelligent Data Analysis, vol. 1, no. 3, pp.131-156, 1997
[6] W. Duch, T. Winiarski, J. Biesiada, J, and A. Kachel, "Feature Ranking, Selection and Discretization," Int. Conf. on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP), pp. 251 – 254, 2003.
[7] Isabella Guyon and Andre Elisseeff, "An Introduction to Variable and Feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157 – 1182, 2003.
[8] M. K. Cope, H. H. Baker, R. Fisk, J. N. Gorby and R. W. Foster, "Prediction of Student Performance on the Comprehensive Osteopathic Medical Examination Level Based on Admission Data and Course Performance," Journal of the American Osteopathic Association, vol. 101, no. 2, pp. 84 – 90, 2001.
[9] Dash M, Liu H (1997) Feature selection for classification. International Journal of Intelligent Data Analysis 1: 131-156.
[10] Ellen pitt, Richi nayak,"The use of various data mining and feature selection methods in the analysis of a population survey dataset", Australlian computer socity inc 2007.
[11] S.Vanaja, K.Ramesh kumar "Analysis of Feature Selection Algorithms on Classification: A Survey" International Journal of Computer Applications (0975 – 8887),Volume 96– No.17, June 2014.
[12] L.Latha, T.deepa,"Feature selection methods and algorithms", International journal on computer science and engineering, Vol. 3 No. 5 May 2011.
[13] Veerabhadrappa,Lalitha Rangarajan, Multi-Level Dimensionality Reduction Methods Using Feature Selection and Feature Extraction, International Journal of Artificial Intelligence & Applications, Volume 1, Number 4, 2010, pp. 54-68.
[14] Sandya H. B., Hemanth Kumar P. , Himanshi Bhudiraja, Susham K. Rao, Fuzzy Rule Based Feature Extraction and Classification, International Journal of Soft Computing and Engineering, Volume 3, Issue 2, 2013, pp. 42-47.
[15] Tomasz Kajdanowicz, Przemysław Kazienko, Piotr Doskocz, Label-Dependent Feature Extraction in Social Networks for Node Classification, Lecture notes in computer science (Springer), volume 6430, 2010, pp.89-102.
[16] V.P.Gladis Pushpa Rathi,Dr.S.Palani,A Novel Approach For Feature Extraction And Selection on MRI Images for Brain Tumor Classification, International Journal of Computer Science and Information Technology, Volume 2, Issue 1, 2012, pp. 225-234.
[17] P.Kavipriya,Dr.Karthikeyan." Case Study: On Improving Student Performance Prediction in Education Systems using Enhanced Data Mining Techniques. "- International Journal of Advanced Research in Computer Science and Software Engineering,- Volume 7, Issue 5, May 2017
[18] P. Kavipriya –"A Review on Predicting Students' Academic Performance Earlier, Using Data Mining Techniques"- International Journal of Advanced Research in Computer Science and Software Engineering,- Volume 6, Issue 12, December 2016