Enhanced Feature Selection Algorithms to Overcome High Dimensionality Problem

V. Arul Kumar
Assistant Professor, School of Computer Science and Applications, REVA University, Bangalore
arulkumarvenugopal@gmail.com

Abstract-Classification is one of the important techniques of data mining. In the classification task, features play a vital role. Therefore, selecting the relevant features becomes an essential task. Though many feature selection algorithms are available many research works are being carried out to improve the classification accuracy. In this paper, a new methodology is proposed with three different feature selection algorithms to improve the classification accuracy by selecting the relevant features.

Index Terms: Data Mining, Feature Selection, Filter Approach, ArKFS, MFSPFA, MFCPFA

1. INTRODUCTION

The last decade had witnessed a tremendous growth in the field of data mining in terms of both, a number of instances and number of features. This growth causes serious problems to many existing data mining algorithms. Data mining applications consist of the high dimensionality of data which contain many inappropriate features. Feature selection is an important and frequently used technique in data mining for dimensionality reduction by removing irrelevant, redundant and noisy features. It brings the immediate effects of speeding up of data mining algorithms, by selecting the relevant features and improving classification accuracy. The past literature showed that various research works were carried out to select the most relevant features and to improve the classification accuracy, but still the problems persist. Hence, three algorithms to overcome the problem.

2. FEATURE SELECTION

Feature Selection is an important technique in many fields, such as Data Mining, Machine Learning and Pattern Recognition. It is a process of selecting relevant features from the original dataset [1]. Feature Selection technique is broadly classified into three, they are, Filter, Wrapper, and Hybrid Approaches. Filter approach selects the relevant features by looking at the intrinsic characteristic of data without the involvement of any learning algorithms [2]. The predetermined learning algorithms [3]. The Wrapper approach selects the relevant features using The Hybrid approach selects the relevant features by combining the characteristics of filter and wrapper approach [4].

3. THE K-NEAREST NEIGHBOR ALGORITHM

In data mining, k-Nearest Neighbour algorithm [5] is used to classify a new object based on attributes and training samples. This algorithm classifies the objects based on closest training examples in the feature space. It is an example of instance-based learning, in which the training data set is stored so that a classification for a new unclassified data may be found simply by comparing it to the most similar records in the training set. First, the algorithm determines the k value, then the distance between test data and all the training data is computed using some distance function d(x, y) to find the nearest neighbor list. Finally, based on the nearest neighbor list, the test data is classified by a majority vote of its neighbors.

4.1. *Distance Function*

Let us consider the two input vectors *X* and *Y*, where $X = \{x_1\}$ and $Y = \{y_1\}$. The distance between two vectors is computed by the following eq. (1)

$$d_{Euclidean}(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \qquad \text{Eq. (1)}$$

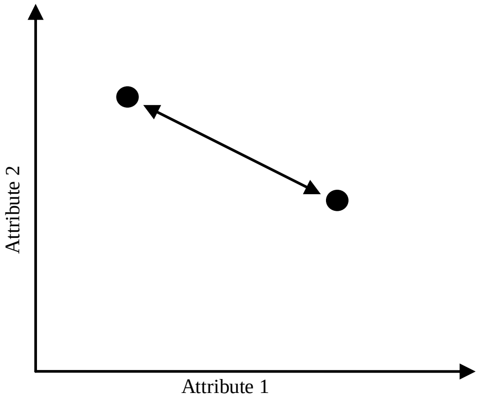

Fig. 1 Euclidean Distance

The steps of k-NN algorithm [5], are given below and the graphical examples are depicted in Figures 2, 3 and 4. In these figures, two classes are used. The square represents the data of class A, circle represents the data of class B and the X represents the data to be tested.

Step 1: Determine parameter k = number of nearest neighbours

Step 2: Calculate the distance between the test data and all the training data

Step 3: Sort the distance and determine the nearest neighbors based on the kth minimum distance

Step 4: Gather the category/classes of the nearest neighbors.

Step 5: Use simple majority of the category of the nearest neighbours to determine the class of the test data.

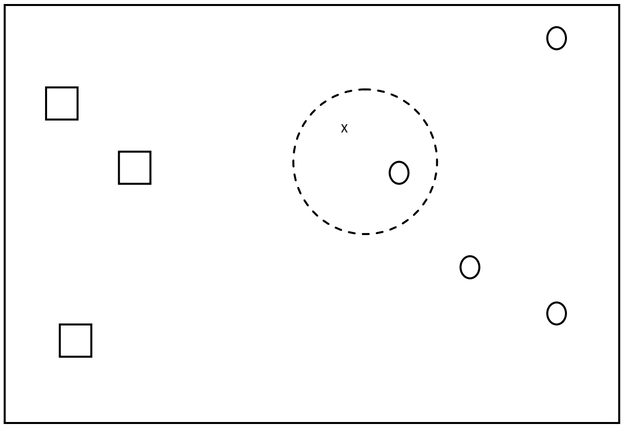

Fig. 2 *k*-Nearest Neighbor Classification (where k=1)

Fig. 2 shows the result of the k-NN algorithm with the k-value 1. The X data is very close to the data which is present in the class B. So, the tested data X is classified as class B.

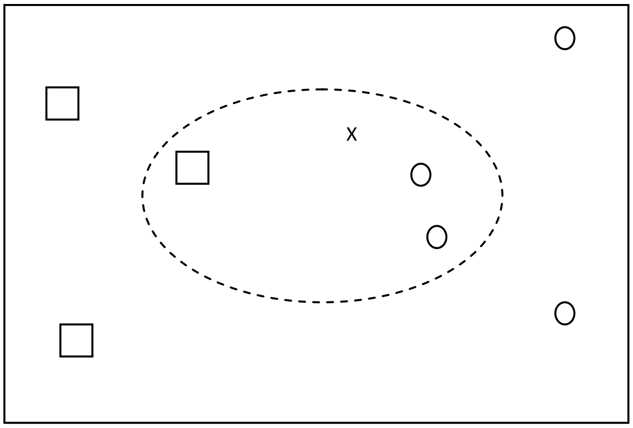

Fig. 3 *k*-Nearest Neighbor Classification (where k=3)

In Fig. 3, the k-value is considered as 3. In this case, the X data finds the three nearest data in which the two data belong to the class B and one data belongs to the class A. The test data X find two nearest data in the class B. Hence, the X data is classified as B.

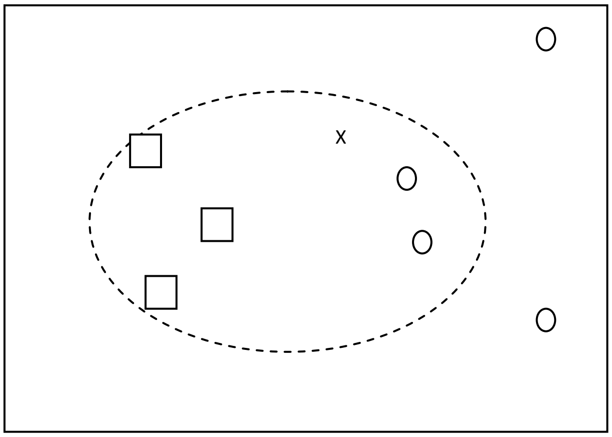

Fig. 4 *k*-Nearest Neighbour Classification (where k=5)

In Fig. 4, the k-value is assigned to 5, the data X finds the five nearest data. The result shows that the X finds the three nearest data in class A and two nearest data in class B. So, the test data X is classified as A because the most nearest neighbor present is in the class A.

4. THE NAIVE BAYESIAN ALGORITHM

The Naïve Bayesian [6] is a simple probabilistic classifier based on Bayes' theorem with strong (naïve) independence assumption. It is used for estimating the probability of each class value during classification and prediction. In simple terms, a Naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. The conditional probability of the selected class is computed by the following eq. (2).

$$P(C_i|X_{test}) = \frac{P(C_i)}{P(X_{test})} \prod_{m=1}^{n} P(x_m \mid C_i) \qquad \text{Eq. (2)}$$

where,

C_i is the i^{th} class,

X_{test} is a test data,

x_m is the value of the m^{th} feature in the X_{test} data.

4.1. *Bayes' Theorem*

The Bayes' theorem [7] is developed by Reverend Thomas Bayes. It is used to estimate the likelihood of a property given the set of data as evidence or input.

$$P(A|B) = \frac{P(B|A) . P(A)}{\sum_i P(B|A_i) . P(A_i)} \qquad \text{Eq. (3)}$$

where,

A is a hypothesis,

B is an observable event,

P (A|B) is the posterior probability,

P (A) is prior probability associated with hypothesis *A*,

P (B|A$_i$) is the conditional probability.

The other way to write the Bayes rule is

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad \text{Eq. (4)}$$

The algorithm classifies the data in two steps; Training step and Prediction step.

- *Training Step :*

  Using the training samples, the method estimates the parameters of a probability distribution, by assuming features that exist in the independent class.

- *Prediction Step*

  For any unseen test sample, the method computes the posterior probability of that sample belonging to each class. The method then classifies the test sample according to the largest posterior probability.

5. THE J48 ALGORITHM

In data mining, a decision tree is one of the most widely used inductive learning methods to classify the given input data. It is originally implemented in Decision Theory (DT) and Statistics. The DT models are used to examine the data and induce the tree and their rule is used to make predictions. The main goal of the DT is to classify the data into discrete groups which make a strong separation in the values of the dependent variable [8]. The Decision Tree includes various types of algorithms such as ID3 [9], C4.5 [10], CART [11], J48 [12] and C5 [13]. The J48 method is an algorithm for decision tree generation and an extension of ID3

6. DATASET

Nine different datasets are used for validating the proposed algorithm. The dataset is obtained from the UCI Machine Learning Repository. The characteristics of the dataset are depicted in Table 1. The motivation for choosing these datasets is due to their popularity. Many researchers use them for their evaluation. They are regarded as a benchmark among Data Mining researching community.

6.1. *UCI Machine Learning Repository*

The UCI (University of California, Irvine) Machine Learning Repository was created by David Aha and fellow graduate students at University of California, Irvine in 1987. The repository consists of different kinds of databases which were used by the machine learning community to validate the machine learning algorithms. But nowadays, this repository is accessed by students, educators, and researchers for data mining. This repository contains 156 datasets for classification analysis, 22 datasets for regression analysis, 13 datasets for clustering analysis and 46 datasets for other analysis. The UCI repository is widely used in the field of computer science especially in data mining [14]. The researcher used nine datasets from the UCI repository and a brief description of the same is given below.

Table 1: Characteristics of the Dataset

Dataset Name	**Total Number of Instances**	**Number of Features**	**Missing Values**
Credit Approval	690	16	Yes
Ionosphere	351	35	No
Annealing	798	39	Yes
Covertype	581012	55	No
Musk v1	446	169	No
Arrhythmia	452	280	Yes
Madelon	4400	500	No
Isolet	7797	617	No
Multiple features	2000	649	No

6.1.1. *Credit Approval Dataset*

The Credit Approval dataset contains 15 features with 690 instances. The dataset contains missing value. The missing values present in the dataset are replaced with a mean value of the feature.

6.1.2. *Ionosphere dataset*

It is a radar dataset collected by a system in Goose Bay, Labrador. The dataset is divided into two classes, namely, Good and Bad. The Good radar shows some types of structures in the ionosphere, whereas Bad radar does not show any of the structures. The dataset contains 351 instances, 34 features (or attributes) with no missing values.

6.1.3. *Annealing Dataset*

The dataset was originally created by David Sterling and Wray Buntine. The annealing dataset consists of 38 features (or attributes), with 798 instances. The dataset contains missing values.

6.1.4. *Covertype Dataset*

The Covertype dataset was determined from US Forest Service (USFS) Region 2 Resource Information System (RIS). The dataset consists of 54 features with 581012 instances. There are no missing values in this dataset.

6.1.5. *Musk V1 dataset*

The dataset was created by AI Group at Arris Pharmaceutical Corporation, in South San Francisco, CA. The dataset consists of 166 features with 476 instances. The instances are categorized into two, namely, musk and non-musk. The musk category contains 207 instances and the non-musk category contains 269 instances. This dataset doesn’t contain any missing values.

6.1.6. *Arrhythmia Dataset*

The Arrhythmia Dataset was originally owned by Altay Guvenir, Burak Acar and Haldun Muderrisoglu. The dataset contains 279 features (or attributes) with 452 instances. The dataset is classified into three classes, they are, “N” with 245 instances, “A” with 185 instances, “U” with 22 instances. This dataset contains missing values.

6.1.7. *Madelon Dataset*

The dataset was generated by the conference organizer (Neural Information Processing Systems Conference). The dataset consists of 500 features with 4400 instances. The features are continuous and normalized, without any missing values in the dataset.

6.1.8. *Isolet Dataset*

The Isolet Dataset is created by Ron Cole and Mark Fanty, Oregon Graduate Institute, Beaverton. The dataset consists of 617 features with 7797 instances. Missing values exist in this dataset.

6.1.9. *Multiple features Dataset*

The dataset was created by Robert P.W. Duin, Delft University of Technology, Netherland. This dataset consists of features of handwritten numerals extracted from a collection of Dutch utility maps. The dataset contains 649 features, 2000 instances. No values are found to be missing in this dataset.

7. PROPOSED ALGORITHMS

In this research article, three different algorithms were proposed to select the required features in the given dataset. The three newly proposed feature selection algorithms, namely, Modified Fisher Score Principal Feature Analysis (MFSPFA), Modified Fisher Criterion Principal Feature Analysis (MFCPFA), ArK Feature Selection (ArKFS). These algorithms are used to select the relevant features from the original features available in the dataset.

7.1. *MFSPFA Algorithm*

The proposed MFSPFA algorithm uses a statistical measure to select the relevant features. The selected features are then applied to three different classifiers to find the classification accuracy. To show enhancement of classification accuracy of the proposed algorithm, nine different datasets, which are obtained from the UCI (University of California, Irvine) Machine Learning Repository are used. The detailed description about UCI repository and the dataset are given in section 3.3.

7.2. *MFCPFA Algorithm*

The existing algorithms take larger dimension space to select the relevant features. So, a new algorithm MFCPFA is proposed to reduce the dimension space and to select the relevant features. To prove the efficiency of the proposed MFCPFA algorithm, it is applied to nine different datasets.

7.3. *ArKFS Algorithm*

The ArKFS algorithm is proposed to overcome the Small Sample Size (SSS) problem. The ArKFS selects the relevant features with the small number of training samples. The efficiency of the ArKFS is analyzed in terms of classification accuracy. Three classifiers are used to compute the classification accuracy of the proposed algorithm. For this, nine different datasets are used. The obtained classification accuracy results are finally compared with the existing feature selection algorithms.

8. CONCLUSION

In this paper, three algorithms were presented to overcome the high dimensionality problem in the field of data mining and to improve the classification accuracy of the data mining algorithms. In this methodology, three feature selection algorithms have been proposed to select relevant features. The enhanced classification accuracy of the proposed features selection algorithms is derived using the three different classification algorithms. The results of the proposed algorithms show the improved classification accuracy.

REFERENCES

[1] Pushpalata Pujari, Jyoti Bala Gupta, Improving Classification Accuracy by Using Feature Selection and Ensemble Model, International Journal of Soft Computing and Engineering, Volume 2, Issue 2, 2017, 380-386.

[2] Jinjie Huang, Yunze Cai, and Xiaoming Xu, A Filter Approach to Feature Selection Based on Mutual Information, Proceedings of the IEEE International Conference on Cognitive Informatics, Volume 1, 2006, pp. 84-89.

[3] Huilin Zhou, Jianbin Wu, Yuhao Wang, and Mao Tian, Wrapper Approach for Feature Subset Selection using GA, Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems, 2007, pp. 181-191.

[4] Li-Yeh Chuang, Kuo-Chuan Wu, and Cheng-Hong Yang, Hybrid Feature Selection Method using Gene Expression Data, Proceedings of the IEEE Conference on Soft Computing in Industrial Applications, 2008, pp. 199-204.

[5] Antonio Mucherino, Petraq J. Papajorgji, and Panos M. Pardalos, k-Nearest Neighbor Classification, Journal of Data Mining in Agriculture, Volume 34, 2009, pp. 83-106.

[6] Min-Ling Zhang, José M. Peña, and Victor Robles, Feature Selection for Multi-Label Naïve Bayes Classification, International Journal of Information Science, Volume 179, Issue 19, 2016, pp. 3218-3229.

[7] Guangzhi Qu, Hui Zhang, and Hartrick, C.T., Multi-label Classification with Bayes' Theorem, Proceedings of the International Conference on Biomedical Engineering and Informatics, Volume 4, 2011, pp. 2281-2285.

[8] Alok Sharma, and Kuldip K. Paliwal, Rotational Linear Discriminant Analysis Technique for Dimensionality Reduction, IEEE Transactions on Knowledge and Data Engineering, Volume 20, Issue 10, 2017, pp. 1336-1347.

[9] Liu Yuxun, and Xie Niuniu, Improved ID3 Algorithm, Proceedings of the IEEE International Conference on Computer Science and Information Technology, Volume 8, 2010, pp. 465-468.

[10] Li Rui, Wei Xian-mei, and Yu Xue-wei, The Improvement of C4.5 Algorithm and Case Study, Proceedings of the 2nd International Symposium on Computational Intelligence and Design, Volume 2, 2016, pp. 190-192.

[11] Deepali Saini, and Anand Rajavat, Performance of Decision Tree Algorithms in Knowledge Based System, International Journal of Computer Science & Information Technology Volume 1, Issue 10, 2011, pp. 734-743.

[12] Yogendra Kumar Jain, and Upendra, An Efficient Intrusion Detection Based on Decision Tree Classifier Using Feature Reduction, International Journal of Scientific and Research Publications, Volume 2, Issue 1, 2012, pp. 1-6.

[13] Meng Wang, Kun Gao, Li-jing Wang, and Xiang-hu Miu, A Novel Hyperspectral Classification Method Based on C5.0 Decision Tree of Multiple Combined Classifiers, Proceedings of the 4th International Conference on Computational and Information Sciences, 2012, pp. 373 – 376.

[14] UCI Machine Learning Repository, (http://archive.ics.uci.edu/ml /datasets.html/dated: 08/08/2012).