Feature Selection Techniques using for High Dimensional Data in Machine Learning

M.sujatha [1], Dr. G. Lavanya Devi [2]
Department of Computer Science and Systems Engineering
AU College of Engineering, Andhra University
Visakhapatnam, Andhra Pradesh, India

Abstract

In machine learning the classification task is normally known as supervised learning. In supervised learning present a specified set of classes and objects are labeled with the correct class. The goal is to generalize from the training objects that will make possible novel objects to be identified as belonging to one of the classes. Evaluating the performance of learning algorithms is a fundamental aspect of machine learning. The main objective of this paper is to study the classification accuracy using feature selection with machine learning algorithms. The dimensionality of the data is reduced by implementing Feature selection and accuracy of a learning algorithm improved. The filtered feature space that is, the condensed feature space provided by a filter. The advantage of feature selection for learning can include a reduction in the amount of data required to achieve learning, improved predictive accuracy, learned knowledge that is more compact and easily comprehend, and reduced execution time.

1. Introduction

Data mining takes the use of data analysis tools to discover previously unknown, valid patterns and relationships from large amounts of data stored in databases, data warehouses, or other information repositories. Data mining has two methods. The first method attempts to produce an overall summary of a set of data to identify and express relevant features. The second method, pattern detection, tries to find small unusual patterns of behaviour. The data mining analysis tasks in generally grouped into data summarization, classification, prediction, segmentation, dependency analysis.

Machine learning use tools by which large quantities of data can be automatically analyzed. The important concept of machine learning is feature selection. Feature selection for clustering or classification tasks can carried out on the basis of correlation between relevant features, and feature selection process can be helpful to a variety of common machine learning algorithms. The feature selector is simple and fast to execute. It eliminates irrelevant and redundant data and, in several cases improves the performance of learning algorithms.

2. Feature Selection

Feature selection is common in machine learning. Feature selection is termed as feature subset selection, variable selection or attributes reduction. Feature selection is the process of selecting the input attributes of a data set that most closely define a particular outcome. The basic three steps of this system are:

- In first step the irrelevant features are removed.
- After that the redundant features are removed.
- And finally a feature selection algorithm is applied to the remaining features.

In this approach each step is working as a filter that reduces the number of candidate features, until finally only a small subset remains.

The first filter removes irrelevant features using a Relief algorithm, which gives relevance values to features from training samples as feature space. There are several modifications to Relief to generalize it for continuous features and to make it more robust in the presence of noise. This system adopts Kononenko's

modifications, and modifies Relief again to remove a bias against non-monotonic features, as described in [1].

Figure 1. Filter feature selection

Within this feature selection system, Relief is used as a relevance filter. Therefore it threshold the relevance values, to divide the feature set into relevant and irrelevant features. This can be done either by threshold the relevance value directly, or by selecting the highest n values and discarding the remaining features. In either case, relief does not detect redundancy, so the remaining feature set still contains redundant features. The second step is a redundancy filter that uses the K-means algorithm [2]

To cluster features according to how well they correlate to each other. When feature clusters are discovered, only the feature with the highest Relief score is kept; the other features in the cluster are removed from the feature set. This is an unusual application of K-means clustering, in that features are clustered (instead of samples), and correlation is used as the distance measure. The third and final filter is a combinatorial feature selection algorithm. The following algorithms are used for to perform the above mentioned operations.

2.1. ReliefF Algorithm

ReliefF (RFF) is an extension to relief algorithm. And proposed by Kira and Rendell in 1994. Relief is an easy to use, fast and accurate algorithm even with dependent features and noisy data [3]. Relief algorithm evaluates the ability of an attribute apart from similar instances. The process of ranking the features in relief follows three basic steps:

- Compute the nearest miss and nearest hit.
- Compute the weight of a feature.
- Compute a ranked list of features or the top k features according to a given threshold.

The basic idea of ReliefF is to illustrate instances at random, calculate their nearest neighbors, and modify a feature weighing vector to give more weight to features that differentiate the instance from neighbors of different classes [4]. It is also improved to deal with noisy data and can be used for regression problems.

2.2 OneR Attribute Evaluation (One R)

Rule based algorithms generating classification rules is to use decision trees. The drawback of using a decision tree is because it is complex and incomprehensible [5]. A classification rule can be defined as r = (p, q) where p is a precondition which performs a sequences of tests that can be estimated as true or false and q is a class that can be appropriate to instances covered by rule r. A general rule of a rule based algorithm tries to cover all instances fit into a class q in a given time. OneR is the simplest approach to finding a classification rule as it creates one level decision tree. OneR constructs rules and tests a individual attribute at a time and branch for every value of that attribute. For every branch, the class with the best classification is the one taking place often in the training data.

2.3. Information Gain (IG)

Information Gain is primary concept of entropy. The expected value of information gain is the mutual information of target variable (X) and independent variable (A). It is the reduction in entropy of target variable (X) achieved by learning the state of independent variable (A) [6]. The major problem of using information gain is to choose attributes with large numbers of discrete values over attributes with fewer values even though the later is more informative. In order to estimate information gain, consider an attribute X and a class attribute Y. The information gain of a given attribute X with respect to class attribute Y is the reduction in uncertainty about the value of Y when the value of X is known. The value of Y is measured by its entropy, H(Y) [6]. The uncertainty about Y, given the value of X is given by the conditional probability of Y given X, H (Y|X).

$$I(Y;X) = H(Y) - H(Y/X) \qquad 2.1$$

where Y and X are discrete variables that take values in $\{y_1.....y_k\}$ and $\{x_1....x_k\}$ then the entropy of Y is

given by:

$$H(Y) = -\sum_{i=1}^{i=k} P\left(Y = y_j\right) \log_2 P\left(Y = y_j\right) \quad 2.2$$

The conditional entropy of Y given X is

$$H(Y/X) = -\sum_{j=1}^{l} P\left(X = x_j\right) H\left(Y/X = x_j\right) \quad 2.3$$

Alternatively the information gain is given by:

$$I(Y;X) = H(X) + H(Y) - H(X,Y) \quad 2.4$$

Where H(X,Y) is the joint entropy of X and Y:

$$H(X,Y) = -\sum_{i=1}^{k}\sum_{j=1}^{l} P\left(X = x_j, Y = y_j\right) \log_2 P\left(X = x_j, Y = y_j\right) \quad 2.5$$

When the predictive variable X is not discrete but continuous, the information gain of X with class attribute Y is computed by considering all possible binary attributes, X_Θ, that arise from X when we choose a threshold Θ on X [6]. Θ takes values from all the values of X. Then the information gain is simply:

$$I(Y;X) = \arg\max_{X_\Theta}(Y, X_\Theta) \quad 2.6$$

2.4. Gain Ratio (GR)

The information gain selects attributes having a large number of possible values over attributes with fewer values even though the later is more informative [7]. For example consider an attribute proceeds as a unique identifier, such as a customer- id in a bank database. A split on customer- id generates large number of partitions; as each record in the database has a unique value for customer-id. So the information required to classify database with this partitioning would be $\text{Info}_{\text{customer-id}}(D) = 0$. Clearly, such a partition is useless for Classification. C4.5, a successor of ID3 [8], uses an extension to information gain known as gain ratio (GR), which attempts to overcome the bias. Let D be a set consisting of d data samples with n distinct classes. The expected information needed to classify a given sample is given by

$$I(D) = \sum^{n} p_i \log_2(p_i) \quad 2.7$$

where p_i is the probability that an arbitrary sample belongs to class C_i. Let attribute A have v distinct values. Let d_{ij} be number of samples of class C_i in a subset D_j. D_j contains those samples in D that have value a_j of A. The entropy based on partitioning into subsets by A, is given by

$$E(A) = -\sum_{i=1}^{n} I(D) \frac{\left(d_{1i} + d_{2i} + \ldots\ldots\ldots + d_{mi}\right)}{d} \quad 2.8$$

The encoding information that would be gained by branching on A is

$$\text{Gain}(A) = I(A) - E(A) \quad 2.9$$

C4.5 applies a kind of normalization to information gain using a "split information" value defined analogously with Info (D) as

$$\textit{SplitInfo}\ \ A(D) = -\sum_{j=1}^{v} \left(\frac{|D_j|}{|D|}\right) \log \left(\left|\frac{D_j}{D}\right|\right) \quad 2.10$$

This value represents the information computed by splitting the dataset D, into *v* partitions, corresponding to the *v* outcomes of a test on attribute A[7]. For each possible outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in D. The gain ratio is defined as

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)} \quad 2.11$$

The attribute with maximum gain ratio is selected as the splitting attribute.

2.5. Symmetrical Uncertainty

Symmetric uncertainty evaluates the degree of association between discrete features. It is derived from entropy [17].Correlation based feature selection is the base for symmetrical uncertainty (SU). Correlation based feature selection estimates the merit of a feature in a subset using a hypothesis – "Good feature subsets contain features highly correlated with the class, yet uncorrelated to each other" [18]. It is a symmetric measure and can be used to compute feature-feature correlation

$$SU = 2.0 \times \frac{H(X) + H(Y) - H(X,Y)}{H(Y) + H(X)} \qquad 2.12$$

Symmetrical uncertainty is computed by the above equation. H(X) and H(Y) represent the entropy of features X and Y. The value of symmetrical uncertainty lies between 0 and 1. The value of 1 specifies that one variable (either X or Y) entirely predicts the other variable [18]. The value of 0 indicates the both variables are totally independent.

3. Classification

Classification is a data mining technique used to calculate group membership for data instances. It is one of the important techniques in data mining and is used in various applications such as pattern recognition, customer relationship management, disease diagnosis and targeted marketing [9]. The various classifiers of data mining are

3.1. K – Nearest Neighbor

In pattern recognition, the *k*-nearest neighbor algorithm (*k*-NN) is a non-parametric technique for classifying objects found on closest training instances in the feature space. *k*-NN is a kind of instance-based learning, or lazy learning where the function is only approximated locally and all calculation is delayed until classification. The *k*-nearest neighbor algorithm is the simplest of all machine learning algorithms: an object is classified based on majority vote of its neighbors, with the object being allocated to the class most common amongst its *k* nearest neighbors (*k* is a positive integer, usually small). If *k* = 1, then the object is simply allocated to the class of that single nearest neighbor.

A data sample in KNN is classified on the basis of a selected number of k nearest neighbors [10]. The KNN assumptions are

- the data is in a feature space, so they have the concept of distance. Euclidean distance can be used to compute distance between vectors.
- Each training vector is associated with set of vectors and class label.
- K decides how many neighbors influence the classification.

3.2. Naïve Bayes

A Naïve Bayes (NB) classifier is a simple probabilistic classifier based on Bayes theorem where every feature is assumed to be class-conditionally independent [11]. In naïve bayes learning, each instance is explained by a set of features and obtains a class value from a predefined set of values. Classification of instances finds complex when the dataset contains a large number of features and classes because it takes huge numbers of observations to approximate the probabilities [11]. When a feature is class-conditionally independent, then the variable value on a given class is independent of those values of other variables.

3.3. Support Vector Machine

A support vector machine (SVM) is a hyperplane that separates two different sets of samples with maximum distance of hyperplane to nearest samples from both sets [12]. The formula for the output of a linear SVM is

$$u = \vec{w} \cdot \vec{x} - b \qquad 3.1$$

in this equation w is the normal vector to the hyperplane and x is the input vector. The nearest points lie on the planes $u = +1$. The distance d is

$$d = \frac{1}{\|W\|_2} \qquad 3.2$$

The maximum distance d can be expressed using optimization problem

$$\min_{\vec{w},b} \frac{1}{2}\left\|\vec{w}\right\|^2 \; subject \text{ to } y_i\left(\vec{w}.\vec{x}_i - b\right) \geq 1, \qquad 3.3$$

where x_i is the i^{th} training sample and y_i is the correct output of the SVM for the i^{th} training sample. The value y_i is +1 for the positive samples and -1 for the negative samples.

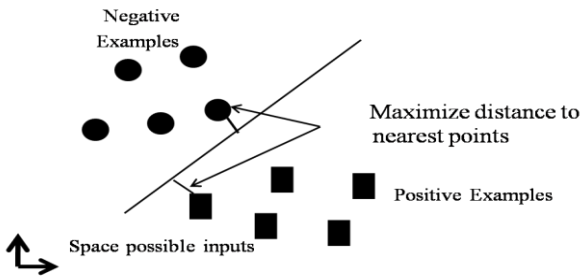

Figure 2. Support Vector Machine

The Sequential Minimal Optimization (SMO) is an algorithm that answers quadratic programming (QP) problem which appears in support vector machine without i relating extra matrix space [12]. SMO divide the complete QP problem into smallest possible QP sub-problems at every step using Osuna's theorem. At each step, SMO finds optimum value of the two Lagrange multipliers and updates the SVM to reflect the new optimum values [12].

3.4. Random Forest

Random forests (RF) are recursive partitioning which merges a collection of trees called an ensemble. Random forests [13] are a group of identically distributed trees whose class value is find by a variant on majority vote. The classifier consists of a collection of tree like classifiers which uses a large number of decision trees, all of which are trained to attempt the same problem.

3.5. C4.5

C4.5 is a variant and extension of an ID3 decision tree algorithm [14].It is based on the concept of a decision tree. A decision tree is a hierarchical collection of rules that describe how to divide a large collection of data into groups based on the regularities of the data [15]. It is a graphical structure used for regression, classification, prediction function and clustering. The objective of a decision tree is to construct an accurate classifier and build up understandable patterns that can be understand as interesting knowledge.

C4.5 contains three types of tests: standard test, complex test and binary test. All tests are based on a discrete attribute. These tests are evaluated using gain ratio.

4. Evaluation Criteria

A classification algorithm is a function that given a set of training samples and their classes constructs a classifier. A classifier is a function that given an instance assigns it to one of the predefined classes. There are a variety of classifiers that have been developed. The main question that arises in the development and application of theses algorithms is about the accuracy of the classifiers they produce. We will be using AUC as evaluation criteria in our thesis which will be discussed in this chapter. AUC is an acronym for Area under Receiver Operating Characteristic Curve [16]. An ROC graph is a technique for visualizing, organizing and selecting classifiers based on their performance. Given a classifier and an instance, there are possible outcomes for the instance. If the instance is positive and it is classified as positive, then it is counted as true positive (TP). If it is classified as negative, then it is counted as false negative (FN). If the instance is negative and it is counted as false positive (FP). If we consider a whole training set we can build a confusion matrix from this methodology [16].

$$confusion\ matrix = \begin{bmatrix} TP & \mathrm{FP} \\ FN & \mathrm{TN} \end{bmatrix} \qquad 4.1$$

The diagonal (upper left to lower right) of the confusion matrix represent the correct decisions made and the elements of the diagonal (upper right to lower left) represent the errors. The true positive rate of a classifier can be estimated as

$$\mathrm{TP\,rate} = \frac{\mathrm{Positives\,correctly\ classified}}{\mathrm{Total\,Positives}} \qquad 4.2$$

The false positive rate can be defined as

$$\mathrm{Fp\,rate} = \frac{\mathrm{Negativelyincorrectly\ classified}}{\mathrm{Total\,Negatives}} \qquad 4.3$$

An ROC graph depicts relative trade-offs between true positives and false positives. To find a clear dominating relation between two ROC curves we use AUC which provides a single-number summary for the performance of` learning algorithms.

We have applied the five feature ranking techniques (GR, RFF, SU, OneR, IG, and Ensemble) to the Lung Cancer dataset. We have selected the top k (k=5) feature subsets for the experiments. After the feature selection, we used five learners, KNN, C4.5, NB, RF, and SVM, to build classification models on the datasets with various selected subset of features. The classification models are evaluated in terms of the AUC performance metric. The results of the experiments are displayed in Table1,2 and 3. Each value in the table is determined by the row (ranker) and the column (learner) in which the value is loaded. It also depends on the value of k used for the table. The process of calculating AUC value for a table is performed in three steps:

1. Identify the row and column for which the AUC needs to be calculated. This helps in selecting a ranker and a learner.
2. Ranker is applied to the dataset to get the ranking list. The top k features are selected from the ranking list. The value of k can be determined by checking the table for which the AUC is calculated.
3. Classification model is built using the dataset with selected features from the previous step.

Table 1. AUC values for rankers with top five feature for lung cancer dataset.

Rankers	*KNN*	*C4.5*	*NB*	*SVM*	*RF*	*AVG*
RFF	0.844	0.769	0.831	0.835	0.848	0.825
OneR	0.758	0.757	0.781	0.773	0.764	0.766
IG	0.877	0.743	0.827	0.821	0.835	0.820
SU	0.877	0.744	0.827	0.821	0.835	0.820
GR	0.735	0.749	0.784	0.773	0.739	0.756

Table 2. Average model performances for classifiers using lung cancer dataset

Classifier	AUC
KNN	0.7762
C4.5	0.7823
NB	0.8423
RF	0.7860
SU	0.7633

Table 3. Average model performances for rankers using lung cancer dataset

Ranker	AUC
SU	0.7871
OneR	0.7397
GR	0.7598
RFF	0.7887
IG	0.7834

The Tables 1, 2 and 3 summarize the classification performance in terms of AUC for the five selected rankers and ensemble method with top k features. The Tables also display model performance on base dataset. All these results are mapped into a group of features as shown in Figures 1 experiments can be summarized in terms of size of feature subset, classifiers and rankers in the following tables. Table 1 shows that selecting top 5 features subset generates highest classification accuracy NB has the highest classification accuracy over other classifiers while LR performed worst. Table 1, 2 and 3 shows that ensemble ranker performed best over other rankers in terms of AUC performance metric, while OneR performed worst. We also compared the results from the subset of features with the results from the complete set of features (base dataset). We found that the classification performance is improved even after a significant number of features were removed from the original dataset. This demonstrates that feature selection was successfully applied to the Lung Cancer dataset.

5. Conclusion

In this paper, we have reviewed feature selection and explained the basic concept of different feature selection methods: filter, wrapper and hybrid model. We reviewed four filter based feature ranking techniques and one wrapper based feature ranking technique. They are information gain, gain ratio, symmetrical uncertainty, reliefF and oneR attribute evaluation. We examined classification models that are built using various classification techniques such as naïve bayes, k-nearest neighbor, random forest, and support vector machine, regression and decision trees. We took a brief review of the evaluation criteria used to evaluate the classification models.

Feature selection is to choose a subset of input variables by eliminating features, which are irrelevant or of no predictive information. The RELIEF is an optimization algorithm, which is used to improve the quality of feature selection.

References

[1] Jadhav, S. R., and Kumbargoudar, P., "Multimedia Data Mining in Digital Libraries: Standards and Features".Proceedings of conference Recent advances in Information Science and Technology READIT – 2007, pages 54-59, Organized by Madras Library Association-Kalpakkam Chapter & Scientific information Resource Division, Indira Gandhi Center for Atomic research, Department of Atomic Energy, Kalpakkam, Tamilnadu,India. 12-13 July 2007.

[2] Witten, I.H., Frank, E. (2005) Data Mining: Practical Machine Learning Tools and Techniques: 2nd Edition. San Francisco: Morgan Kaufmann.

[3] H. Wang, T. M. Khoshgoftaar, K. Gao, "Ensemble feature selection technique for software quality classification", Proceedings of the 22nd International Conference on Software Engineering & Knowledge Engineering, Redwood City, San Francisco Bay, CA, USA, July 1 - July 3, 2010, pages 215-220.

[4] Y. Wang, F. Makedon, "Application of ReliefF feature filtering algorithm to selecting informative genes for cancer classification using microarray data", Computational systems bioinformatics conference, 2004 IEEE, pages 497 – 498.

[5]S. Pu latova, "Covering (rule-based) algorithms Lecture Notes in Data Mining", World Scientific publishing Co, 2006, pages 87-97.

[6] W. Altidor, T. M. Khoshgoftaar, J. Van Hulse, A. Napolitano, "Ensemble feature ranking methods for data intensive computing applications", Handbook of data intensive computing, Springer Science + Business media, LLC 2011, pages 349 -376.

[7] Asha G. K, A. S. Manjunath, M. A. Jayaraim, "A Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection",46 International Journal of Information Technology and Knowledge Management, July – December 2012, Volume 2, pages 271 – 277.

[8] J. R. Quinlan, "Induction of decision tress", Machine Learning, Kluwer Academic Publishers, 1986, pages 81-106.

[9] M. W. Kim, J. W. Ryu, "Optimized Fuzzy Classification for Data Mining", 9th International Conference Y. Lee et al. (Eds): DASFAA 2004, LNCS 2973, pages 582 –593.

[10] H. He, W. Graco, X. Yao, "Application of Genetic algorithm and k-nearest neighbor method in medical fraud detection", Simulated Evolution and Learning, 47 Second Asia-Pacific Conference on Simulated Evolution and Learning, SEAL' 98 Canberra, Australia, Springer, November 1998, pages 74 – 81.
[11] M. Narasimha Murty, V. Susheela Devi, "Pattern Recognition: An Algorithmic approach", Springer, Chapter 4, pages 86 -97.
[12] J. C. Platt, "A Fast Algorithm for Training Support Vector Machines", Technical Report MSR-TR-98-14, April 21, 1998, John Platt Microsoft Research.
[13] T. D. Lemmond, B. Y. Chen, A. O. Hatch, W. G. Hanley, "An Extended Study of the Discriminant Random Forest", Chapter 6, "Data Mining: A Special Issue in Annals of Information systems", Springer Science, LLC 2010.
[14] J. R. Quinlan, "Induction of decision tress", Machine Learning,Kluwer Academic Publishers, 1986, pages 81-106.
[15] O. A. Omitaomu, "Lecture Notes in Data Mining", Chapter 4, World Scientific publishing Co, 2006, pages 39 – 51.
[16] T. Fawcett, "ROC Graphs: Notes and Practical Consideration for Researchers", HP Laboratories, March 16, 2004, Kluwer Academic Publishers, pages 1 -38.
[17] Y. Chen, Y. Li, X. Cheng, L. Guo, H. Lipmaa, M. Yung, D. Lin, "Survey and Taxonomy of Feature Selection Algorithms in Intrusion Detection System", Inscrypt 2006, LNCS 4318 Springer-Verlag, Berlin, 2006, pages 153 – 167.
[18] D. Ienco, R. G. Pensa, R. Meo, "Context-based Distance Learning for Categorical Data clustering", IDA 2009, LNCS 5772, Springer, Berlin, 2009, pages 83 – 94.
[19] L. Rokach, B. Chizi, O. Maimon, "Feature selection by combining multiple methods", Advances in Web Intelligence and Data Mining, 2006, pages 295–304.
[20] Y. Saeys, T. Abeel, Y. V. Peer, "Robust feature selection using ensemble feature selection techniques", W. Daelemans et al. (Eds.): ECML PKDD 2008, Part II, LNAI 5212, pages 313–325.
[21] J. T de Souza, N. Japkowicz, S. Matwin, "Stochfs: A framework for combining feature selection outcomes through a stochastic process", Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2005, pages 667–674.
[22] J. O. S. Olsson, D. W. Oard, "Combining feature selectors for text classification", CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management, New York, NY, USA, 2006, pages 798–799.