IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

FEATURE SELECTION AND CLASSIFICATION TECHNIQUES IN DATA MINING

S.Sabeena*, G.Priyadharshini
Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India.

DOI:

ABSTRACT

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Feature selection is one of the important techniques in data mining. It is used for selecting the relevant features and removes the redundant features in dataset. Classification is a technique used for discovering classes of unknown data. Classification task leads to reduction of the dimensionality of feature space, feature selection process is used for selecting large set of features. This paper proposed various feature selection methods.

KEYWORDS: Data mining, Feature selection, Classification Techniques.

INTRODUCTION

Data Mining is the process of extracting large volumes of raw data from hidden knowledge. The health care industry requires the use of data mining techniques as it generates huge and complex volumes of data. The applications of data mining techniques to medical data extract patterns which are useful for diagnosis, prognoses and treatment of diseases. This extraction of patterns allows doctors and hospitals to be more effective and more efficient. The huge volume of data is the barrier in the detection of patterns [1]. Classification task leads to reduction of the dimensionality of feature space, feature selection process is used for selecting large set of features.

The term Knowledge Discovery from data (KDD) refers to the automated process of knowledge discovery from databases. The process of KDD is comprised of many steps namely data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation. Data mining is a step in the whole process of knowledge discovery which can be explained as a process of extracting or mining knowledge from large amounts of data [2]. Data mining is a form of knowledge discovery essential for solving problems in a specific domain. Data mining can also be explained as the non trivial process that automatically collects the useful hidden information from the data and is taken on as forms of rule, concept, pattern and so on [3].

The knowledge excerpted from data mining, allows the user to find interesting patterns and regularities deeply buried in the data to help in the process of decision making. The data mining tasks can be broadly classified in two categories: descriptive and predictive. Descriptive mining tasks defined in the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. According to various goals, the mining task can be mainly classified into four types: class/concept description, association analysis, classification or prediction and clustering analysis [4]. This paper provides a survey of various feature selection techniques and classification techniques used for mining.

DATA PREPROCESSING

Data preprocessing is a data mining technique that involves transforming raw data into an understandable manner. Real world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data needs to be pre processed before applying data mining techniques which is done using following steps: Data Integration – If the data to be mined it derived from different sources data needs to be integrated which involves removing inconsistencies attributes or attribute value names between data sets of different sources. Data Cleaning – This step may involve identifying and correcting errors in the data, filling in missing values, etc. Data Selection – In this method where data relevant to the analysis task are retrieved from the database. Data transformation – This method involves the data transformed or consolidated into forms appropriate for mining by performing aggregation operations for instance [1].

Feature selection

The amount of data has been growing rapidly in recent years, and data mining as a computational process involving methods at the intersection of learning algorithms, statistics, and databases, deals with this huge volume of data, processes and analyzes. The purpose of data mining is to find knowledge from datasets, which is expressed in a comprehensible structure. Moreover, in the presence of many irrelevant and redundant features, data mining methods tend to fit to the data which decrease its generalization. Consequently, a common way to overcome this problem is reducing dimensionality by removing irrelevant and redundant features and selecting a subset of useful features from the input feature set [5].

Feature selection is one of the important and frequently used techniques in data preprocessing for data mining. It brings the immediate effects for applications such as speeding up a data mining algorithm and improving mining performance. Feature selection has been applied to many fields such as text categorization, face recognition, cancer classification, and finance and customer relationship management. Feature selection is a process of selecting a subset of features from a larger set of features, which leads to the reduction of the dimensionality of feature space for a successful classification task. In the feature selection technique the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context [6].

A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets. The choice of evaluation metric heavily influences the algorithm, and these evaluation metrics which distinguish between the three main categories of feature selection algorithms: filters, wrapper, embedded and hybrid methods. Feature selection is of detecting the relevant features and discarding the irrelevant features [7]. Feature selection has several advantages.

- Improving the performance of classification algorithms.
- Data understanding, gaining knowledge about the process and perhaps helping to visualize it.
- Data reduction, limiting storage requirements and perhaps helping in reducing costs.
- Simplicity, possibility of using simpler models and gaining speed.

CLASSIFICATION METHODS

Data mining algorithms can be classified into three different learning approaches: supervised, unsupervised, or semi-supervised. In supervised learning, the algorithm works with a set of examples whose labels are known. The labels can be nominal values in the case of the classification task, or numerical values in the case of the regression task. In unsupervised learning, in contrast, the labels in the dataset are unknown, and the algorithm consistently aims at grouping examples according to the similarity of their attribute values, characterizing a clustering task. Finally, semi-supervised learning is usually used when a small subset of labeled is available, together with a large number of unlabeled examples [8].

The classification task can be seen as a supervised technique where each instance belongs to a class, which is indicated by the value of a special goal attribute or simply the class attribute. The goal attribute can take on categorical values, each of them corresponding to a class. Each example consists of two parts, namely a set of predictor attribute values and a goal attribute value. The former are used to predict the value of the latter. The predictor attributes should be relevant for predicting the class of an instance [9].

In the classification task the set of examples being mined is divided into two mutually exclusive and exhaustive sets, called the training set and the test set. The classification process is correspondingly divided into two phases: training, when a classification model is built from the training set, and testing, when the model is evaluated on the test set. In the training phase the algorithm has access to the values of both predictor attributes and the goal attribute for all examples of the training set, and it uses that information to build a classification model. This model represents classification knowledge essentially, a relationship between predictor attribute values and classes that allows the prediction of the class of an example given its predictor attribute values. For testing, the test set the class values of the examples is not shown. In the testing phase, only after a prediction is made is the algorithm allowed to see the actual class of the just-classified example [10].

One of the major goals of a classification algorithm is to maximize the predictive accuracy obtained by the classification model when classifying examples in the test set unseen during training. The knowledge discovered by a classification algorithm can be expressed in many different ways like Support Vector Machine, Decision trees, Bayesian network, Nearest Neighbour etc.

Support Vector Machine

Many statistical machine learning methods have been employed for data mining subtasks. Based on a strong mathematical foundation, SVMs (Support Vector Machines) have been of the most actively developed classification and regression methodologies due to their salient properties such as margin maximization and systematic nonlinear classification via kernal tricks [11].

Support Vector Machines (SVM) is a powerful state of the art algorithm with strong theoretical foundations based on the Vapnik Chervonenkis theory. SVM has strong regularization properties. Regularization refers to the generalization of the model to new data. The geometrical interpretation of support vector classification is that the algorithm searches for the optimal separating surface, i.e., the hyperplane that is, in a sense, equidistant from the two classes. Support vector classification has two classifications, namely, linearly separable and non-linear decision surfaces [12].

SVMs are set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classification. A special property of SVM is, SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So, SVM called Maximum Margin Classifiers SVM is based on the Structural risk Minimization (SRM). SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed [12].

Two parallel hyperplanes are constructed on each side of the hyperplane that separate the data. The separating hyperplane is the hyperplane that maximize the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be. Consider data points of theform$\{(x_1,y_1),(x_2,y_2),(x_3,y_3),....,(x_n,y_n)\}$. Where y_n = 1 / -1, a constant denoting the class to which that point x_n belongs. n=number of sample. Each x_n is p dimensional real vector. The scaling is important to guard against variable (attributes) with larger variance [14]. To view this Training data, by means of the dividing (or separating) hyperplane, which takes,

w.x + b = 0

where b is scalar and w is p-dimensional Vector. The vector w points perpendicular to the separating hyperplane. Adding the offset parameter b allows us to increase the margin. Absent of b, the hyperplane is forced to pass through the origin, restricting the solution. As with the interest in the maximum margin, in SVM and the parallel hyperplanes. Parallel hyperplanes can be described by equation

w.x + b = 1
w.x + b = -1

If the training data are linearly separable, select these hyperplanes so that there are no points between them and then try to maximize their distance. By geometry, to find the distance between the hyperplane is 2 / │w│. To minimize │w│, to excite data points need to ensure that,

w. xi –b ≥ 1 or w. xi –b ≤ -1

This can be written as,

$yi (w. xi - b) \geq 1 , 1 \leq i \leq n$

SVMs fall into the intersection of two research areas: kernel methods, and large margin classifiers. SVM has been applied to feature selection, time series analysis, reconstruction of a chaotic system, and non-linear principal components. Further advances in these areas are to be expected in the near future. SVMs and related methods are also being increasingly applied to real world data mining [15].

3.2 Decision Tree

Decision tree approach is most useful in classification problem. In this technique, a tree is constructed to form the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple. Decision tree classifier follows the subsequent simple algorithm. To classify a new item, it required to create a decision tree based on the features of the training data. It shows the attribute that discriminates the various instances of data more clearly. This feature tells about the data instances and classify them best is having the highest information gain. If there is no ambiguity among possible features means data instances declining within the group having same value for the target variable. Then conclude the branch and assign to the target value that is obtained [16].

In other cases, look for another feature which gives highest information gain value. If there is no explicit result from the existing information that means it assigns the branch as a target value which possess majority of items. Now the decision tree, follow the order of feature selection as obtained for the tree. By examining all the relevant attributes and their values with those seen in the decision tree model, it can assign or predict the target value of this new instance [17].

3.3 Naïve Bayes

Naïve Bayes classifier works on a simple, but comparatively intuitive concept. It works on the basis of conditional probability. It makes useful for all the attributes which is in the dataset, and analyses them individually as though they are equally important and independent of each other. For example, consider that the training data consists of various animals (say elephants, monkeys and giraffes), and the classifier has to classify any new instance that it encounters. Elephants have attributes like they have a trunk, huge tusks, a short tail, are extremely big, etc. Monkeys are short in size, jump around a lot, and can climbing trees; whereas giraffes are tall, have a long neck and short ears.

When classifying a new instance it will consider each features individually. So if the new instance is an elephant, the Naïve Bayes classifier will not check whether it has a trunk and has huge tusks etc. Rather, it will separately check whether the new instance has a trunk, whether it has tusks, whether it is large, etc. It works under the assumption that one attribute works independently of the other attributes contained by the sample. Though simple in concept, this classifier works well in most data classification problems [18].

3.4 Genetic Algorithm

Genetic algorithms (GAs) are best known for their ability to efficiently search in large spaces which is known as priori. As genetic algorithms are relatively insensitive to noise, it is best choice for feature selection strategy in a robust manner. It is a form of inductive learning strategy, are adaptive search techniques which have demonstrated substantial improvement over a variety of random and local search methods. This is able to use accumulating information about an initially unknown search space in order to bias subsequent search into promising subspaces. Since GAs is basically a domain independent search technique, they are ideal for applications where domain knowledge and theory is difficult or impossible to provide [19].

The main issues in applying GAs to any problem are selecting an appropriate representation and an adequate evaluation function. The main thing in feature selection problem is representing all feature subsets from the existing feature set. The representation will be in binary form. Each feature in the candidate feature set is considered as a binary gene and each individual consists of fixed-length binary string representing some subset of the given feature set. An individual of length *l* corresponds to an *l*-dimensional binary feature vector X, where each bit represents the elimination or inclusion of the associated feature. Then, $x_i = 0$ represents elimination and $x_i = 1$ indicates inclusion of the i^{th} feature. Poor results are discarded, and the good ones preserved. A good solution were combined, and then the whole procedure is repeated [20].

3.5 Nearest Neighbour

A Nearest Neighbor Classifier assumes all instances correspond to points in the n-dimensional space. During learning, all instances are remembered. When a new point is classified, the k-nearest points to the new point are found and are used with a weight for determining the class value of the new point. For the sake of increasing accuracy, greater weights are given to closer points [21].

CONCLUSION

The goal of classification result integration algorithms is to generate more certain, precise and accurate system results. Although or perhaps because many classification methods have been proposed. Classification methods are typically strong in modeling interactions. After analysing all the classification techniques it becomes more flexible to decide a technique for data mining. Mostly decision tree induction is most understandable when compared with the other techniques.

REFERENCES

[1] Jiwawei Han, Micheline Kamber, “Data Mining Concepts and Techniques”, 2000.
[2] Han J, and Kamber M, “Data mining:concepts and techniques”, Morgan Kaufmann, 2006.
[3] Akadi A.E, Ouardighi A.E, Aboutajdine D, “A powerful feature selection approach based on mutual information”, International Journals of Computer Science Netw. Secur, Vol8, pp. 116-121, 2008.
[4] Biesiada J, Duch W, “Feature Selection for High-Dimensional Data: A Pearson Redundancy Based Filter”, Computer Recognition Systems. Springer, pp. 242-249, 2007.
[5] Guyon I, Elisseeff A, “An introduction to variable and feature selection”, J. Mach, Vol 3, 1157-1182, 2003.
[6] He X, Cai D, Niyogi P, “Laplacian score for feature selection”, Adv. Neural Inf. Process Syst, Vol 18, 507-514, 2005.
[7] Aghdam M.H, Ghasem Aghaee N, Basiri M.E, “Text feature selection using ant colony optimization”, Expert Syst. Appl, 6843-6853, 2009.
[8] Himani Bhavsar, Mahesh H.Panchal, “A Review on Support Vector Machine for Data Classification”, International Journal of Advanced Research in Computer Engineering, Vol 1, 2012.
[9] Ferreira A.J, Figueiredo M.A.T, “An approach to feature discretization and selection”, Pattern Recognition, 3048-3060, 2012.
[10] Martínez Sotoca J, Pla F, “Supervised feature selection by clustering using conditional mutual information based distances”., Pattern Recognition, 2068–2081, 2010.
[11] Liu H, Yu L, “Toward integrating feature selection algorithms for classification and clustering”, IEEE Trans, Vol 17, pp. 491-502, 2005.
[12] Mesleh A.M.D, Kanaan G, “Support vector machine text classification system: using ant colony optimization based feature subset selection”, International Conference on Computer Engineering & Systems, pp. 143–148, 2008.
[13] Sugumaran V, Muralidharan V, Ramachandran K.I, “Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing”, 2007.
[14] Himani Bhavsar, Mahesh H.Panchal, “A Review on Support Vector Machine for Data Classification”, International Journal of Advanced Research in Computer Engineering, Vol 1, 2012
[15] Unler A, Murat A, Chinnam R.B, “mr2PSO: a maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification”, Information Science, 4625-4641, 2011.
[16] Priyanka Sharma, “Comparative Analysis of Various Decision Tree Classification Algorithms using WEKA”, International Journal on Recent and Innovation Trends in Computing and Communication, Vol 3, pp. 684–690. (2014).
[17] M. Garofalakis, D. Hyun, R. Rastogi and K. Shim, “Building Decision Trees with Constraints”, Data Mining and Knowledge Discovery, vol. 7, no. 2, pp. 187 – 214, 2003.
[18] U. M. Fayyad, “Branching on attribute values in decision tree generation” AAAI Conference, pp. 601-606, 2001.

[19] Trevino V, Falciani F, “An R package for multivariate variable selection using genetic algorithms”, Bioinformatics, 1154-1156, 2006.

[20] Yang J, Honavar V, “Feature subset selection using a genetic algorithm”, IEEE Transaction”, Vol 13, pp. 44-49, 2008.

[21] Aman Kumar Sharma, Suruchi Sahni, “A Comparative Study of Classification Algorithms for Spam Email Data Analysis”, IJCSE, 1890-1895, 2011.