A Survey on Data mining classification approaches

Anupama Mishra[1], B.B.Gupta[2], Dragan Peraković[3] and Francisco José García Peñalvo[4]

[1]*Swami Rama Himalayan University, India*
[2]*National Institute of Technology, Kurukshetra, Haryana 136119, India & Asia University, Taichung 413, Taiwan & Staffordshire University, Stoke-on-Trent ST4 2DE, UK*
[3]*University of Zagreb, Croatia*
[4]*University of Salamanca, Spain*

Abstract

In this review article, we discuss a number of different classification algorithms used in data mining for unique applications. There are various techniques to analyse the data for continuous and discrete values. Though,in our research paper, we discuss algorithm used for classification and applied for data mining. Basically classification is a technique for categorising data into discrete categories depending on limitations. The Genetic algorithm C4.5, the Naive Bayes algorithm, and others are examples of classification algorithms.

Keywords

Bagging, Naive Bayes, SVM, Random Forest, data mining

1. Introduction

The practise of identifying previously unknown, valid patterns and links in large data sets using advanced data analysis tools is known as data mining. These technologies include statistical models, mathematical algorithms, and machine learning methodologies. There are wide applications of data mining techniques [1, 2]. As a result, data mining includes more than just data collection and maintenance; it also includes analysis and prediction. The classification technique, which can handle a larger range of data than regression, is gaining prominence [3]. Knowledge discovery from datasets is a part of data mining. Data mining tools and methods are applied to extract patterns and features from large amount of data [12], which can then be applied to other datasets[4, 6]. Classification is a process that assigns an object or event to one of the predefined classes in a group. It's based on their characteristics in order to be able to predict their future behavior. Classification methods are used when the data set has already been divided into groups before the classification process begins. The accuracy often depend on the preprocessing of the data which involves data cleaning (missing values, null values, blank values), data integration from multiple sources, data transformation and discretizaion [15].

International Conference on Smart Systems and Advanced Computing (Syscom-2021), December 25–26, 2021
tiwari.anupama@gmail.com (A. Mishra); gupta.brij@gmail.com (B.B.Gupta); dperakovic@fpz.unizg.hr (D. Peraković); fgarcia@usal.es (F. J. G. Peñalvo)

2. Classification Techniques in Data Mining

Classification Techniques are methods of data analysis that can be used to determine the categorization of an individual based on their personal attributes [8, 10]. These techniques help us better understand individuals by grouping them together depending on their lifestyle, habits, and traits. Figure 1 presents the classification algorithms which are generally used for data mining applications.Classification is one of the most commonly used data mining techniques. It can be used for both categorical and numerical attributes. The goal is to predict the class labels of new, unseen observations by using training data consisting of both labeled and unlabeled examples. This method makes use of an algorithm to identify patterns in the training data that are predictive of new observations [25, 18].

2.1. Decision Tree

A decision tree is a class discriminator that iteratively splits the training set until each partition contains only or primarily samples from one class. A split point is a test that describes how data is partitioned in each non-leaf node of the tree based on one or more qualities[13].

2.2. Naive Bayes

Naive bayes is used to work with probabilistic models and majorly used in machine learning [11]. In this model, probability is calculated for each class to determine their categorization, which is then used to forecast the values for a new class. Here, y is a instance of a problem which has to be classified. A vector can represent it by $y = y1, y2, ...yn$ where n represents independent variables, and assigned to instance probabilities $p(cy/(y1, y2, ...yn))$ For each of n possible outcomes or classes cn

$$p(cn|y) = \frac{p(cn)p(y|cn)}{p(Y)}$$

2.3. Rule Based Classification

"If-then-"rules are the classification rules, and the rule is a condition. The rules of individuals are ranked. Rule-based order refers to the order that is based on their quality. Class-based ordering refers to the grouping of rules that belong to the same class. A good rule should be error-free and cover as many scenarios as possible.

2.4. Support Vector Machine

Support Vector Machines (SVM) [23], is a classification technique that can be used to build both classifiers and non-parametric regression models. SVM works by finding an optimal hyperplane that separates objects of different classes in the input space based on their training samples. A new classification method for both linear and non-linear data is the Support Vector Machine. It transforms the original training data into a higher dimension using a non-linear mapping. It searches for the linear optimal separation hyper plane with the additional dimension (i.e. "decision boundary"). A hyper plane with a good non-linear mapping to a high enough

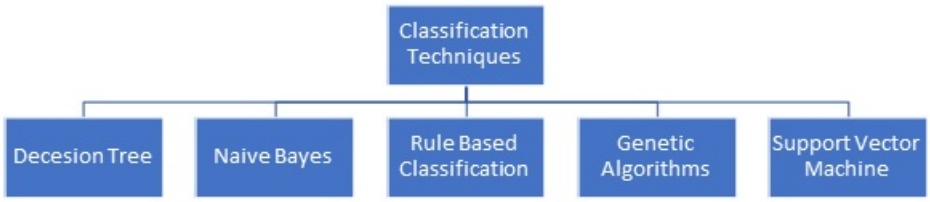

Figure 1: Classification Algorithm

dimension can always divide data into two groups. SVM uses support vectors ("important training tuples") and margins (specified by the support vectors) to discover this hyper plane. SVM is used for classification as well as prediction.

2.5. Genetic Algorithms

In GA, a technique called association rules mining is utilised to uncover indeterminate solutions [9].GA is implemented with a small collection of categorical data. After GA is implemented, high-level prediction rules are produced for the selection of better attribute. The Michigan technique provides a single prediction rule for every individual in the entire population by lowering the cost [7]. The Pittsburgh method [5] is a set of prediction criteria for a whole group of people. We evaluate the overall quality of the rule set rather than the quality of each individual rule when categorising. Rules, like the logical OR implementing and logical AND implementing AND operators, are generalised or specialised based on facts (logical AND).

3. Model Evaluation and Selection

The task of choosing a model is challenging because many models are often equivalent in terms of accuracy, but have different computational complexity. The evaluation and selection of the best model for the particular application depends on the cost-complexity trade-off. One alternative in tackling this task is to carry out an exhaustive search over all possible models, which may be costly in terms of computational time or storage space. Figure 2 depicts the types of methods used for Evaluation and selection of the model [16].

3.1. Hold-Out

A technique used to improve classification accuracy is the holdout validation. It will remove data that was used in training and then split the remaining data into two parts, one for training and one for testing. This prevents over-fitting of the model on the training set.

Hold out validation is a technique that can be used to improve classification accuracy. This method is used by removing any data that was used in training, splitting the remaining data into two groups (one for testing and one for training), and preventing over-fitting of models on the test set by using only new data for testing.

3.2. n-fold Cross Validation

The available data is divided into n distinct subsets of equal size. To train a classifier, use each subset as the training set. The operation is repeated n times, with the given accuracies being the average of the n accuracies. Cross-validation methods such as 10-fold and 5-fold are often utilised. When the available data is small, this strategy is employed.

3.3. Leave-one-out cross validation

if the data volume is small , then this method can be used.Cross-validation is a subset of it. Each cross validation fold contains only one test case, and all of the data's tests are used in training [17]. When there are m examples in the original data set, this is referred to as m-fold cross-validation.

3.4. Validation Set

A validation set is widely used in learning algorithms to estimate parameters. In such instances, the final parameter values are those that provide the highest accuracy on the validation set. Cross validation can also be used to estimate parameters.The data may be divided into the below three sets:
1. Training set
2. Validation set
3. Test set

3.5. Minimum Description Length (MDL)

Missing values are treated by DL as though they were missing at random. Zero vectors are used to replace sparse numerical data while zero vectors are used to replace sparse categorical data [14]. Missing values are considered sparse in nested columns. In the case of MDL, the model's size as well as the reduction in uncertainty that using the model causes does matters [22].

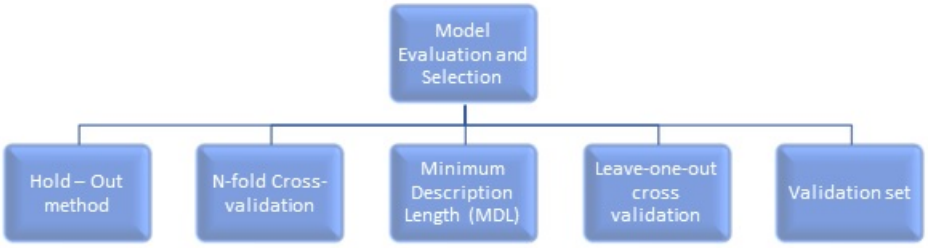

Figure 2: Model Evaluation and Selection

4. Techniques to Improve Classification Accuracy

Classification Accuracy describes how well a model can assign the correct class to a given input. Improvements in Classification Accuracy is important for models that are more accurate and fair as they can reduce the risk of unjustified misclassifications and false alarms.

4.1. Bagging

The most common technique for improving classification accuracy is to use the bagging technique [21]. The bagging technique samples training data across multiple training folds (called bootstrapping) and then uses the resampled data to train the classifier. Boosting is another popular technique which applies weighting to examples that are misclassified by the initial classifier. The bagging technique can significantly improve classification accuracy. During the training phase, each classifier is trained with a subset of the data. This process is called bootstrapping. Once a classifier has been trained for a specific set of data, it's used to classify new data from the same set. In contrast to boosting, wherein each classifier is trained on both positive and negative examples from the input set, bagging trains each classifier on only one type of example at a time. The result is an ensemble of classifiers that combine to create an even better model than any single component could be alone.

4.2. Boosting

It is well-known that boosting algorithms are high-performing classifiers [19]. They are versatile and provide good accuracy when there is a large imbalance in the training data. However, they have a drawback in that they can be computationally expensive for online or near real-time

processing. Boosting is the process of producing a classifier in a successive manner. Each classifier dependents on the preceding based one and concentrates on the errors of the before one. Test sets that have previously been wrongly predicted by classifiers are selected frequently and are weighted properly. The weights of data that are already categorised will be increased. Data that are correctly classified will have their weights reduced. Boosting is a machine learning technique that can be used to improve the classification accuracy of your model. It can be used in many different scenarios, but boosting is most often applied when the goal of the algorithm is to identify what class (or label) an observation belongs to [20].

A big issue with boosting models is that they do not always converge well. This results in algorithms not being able to make accurate predictions. There are many different techniques which can mitigate this issue, one of them being early stopping.

Boosting models work by accumulating error terms, which are then used to adjust weights on different parts of the model or training data. The more data you have for a given class, the larger its weight will be in your model's prediction function and vice versa for other classes.

4.3. Occam's Razor

Occam's razor presents the theory that fits our data and identify unfamiliar objects. It says that if two or more models have same kind of generalisation errors, the simpler model should be preferred over the more complex model. There is a greater probability that sophisticated models will be fitted accidently due to data mistakes [22].

4.4. Random Forest

WEKA is a general-purpose classification and regression tool [24]. For gradient boosting and support vector machine, random forest has a very high accuracy. Random Forest is divided into two types. 1. Regression and classification trees 2. A bootstrap sample is a sample derived from the original dataset with replacement that is the same size as the original dataset.

5. Conclusion

This review discusses numerous data mining classification techniques. each technique has its own set of advantages and disadvantages. Data mining is a broad term that encompasses a variety of approaches for analysing vast amounts of data, including many technology like machine learning and deep learning with the included knowledge of statistics. To accomplish various data analysis tasks, these areas have a huge number of data mining algorithms built in them. Based on the behaviour of the data, the algorithm and evaluation methods can be applied to analyse the data.

References

[1] Mamta (2021) Quick Medical Data Access Using Edge Computing, Insights2Techinfo, pp.1
[2] Sandeep Kumar (2021) Artificial Intelligence and Machine learning for Smart and Secure Healthcare System, Insights2Techinfo, pp.1

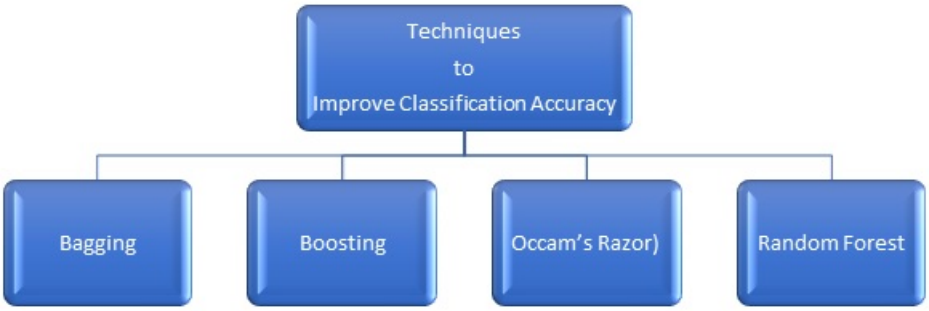

Figure 3: Techniques to Improve Classification Accuracy

[3] Adil, K., Jiang, F., Liu, S., Grigoriev, A., Gupta, B. B., & Rho, S. (2017). Training an agent for fps doom game using visual reinforcement learning and vizdoom. International Journal of Advanced Computer Science and Applications, 8(12).

[4] AlZu'bi, S., Shehab, M., Al-Ayyoub, M., Jararweh, Y., & Gupta, B. (2020). Parallel implementation for 3d medical volume fuzzy segmentation. Pattern Recognition Letters, 130, 312-318.

[5] Witten, I. & Frank, E. (2005), "Data Mining: Practical Machine Learning Tools And Techniques", 2nd Edition, Morgan Francisco, 2005.

[6] Zheng, Z. (2000). Constructing X-Of-N Attributes For Decision Tree Learning. Machine Learning 40: 35–75.

[7] Al-Ayyoub, M., AlZu'bi, S., Jararweh, Y., Shehab, M. A., & Gupta, B. B. (2018). Accelerating 3D medical volume segmentation using GPUs. Multimedia Tools and Applications, 77(4), 4939-4958.

[8] Friedman, N., Geiger, D. &Goldszmidt M. (1997). Bayesian Network Classifiers.Machine Learning 29: 131-163.

[9] Fayyad, U., Piatetsky-Shapiro, G., And Smyth P., "From Data Mining To Nowledge Discovery In Databases," Ai Magazine, American Association For Artificial Intelligence, 1996.

[10] Friedman, N. &Koller, D. (2003). Being Bayesian About Network Structure: A Bayesian Approach To Structure Discovery In Bayesian Networks. Machine Learning 50(1): 95-125.

[11] Quinlan, J.R., C4.5 – Programs For Machine Learning.Morgan Kaufmann Publishers, San Francisco, Ca, 1993.

[12] Bianca V. D.,PhilippeBoula De Mareüil And Martine Adda-Decker, "Identification Of Foreign-Accented French Using Data Mining Techniques, Computer Sciences Laboratory For Mechanics And Engineering Sciences (Limsi)".

[13] Breslow, L. A. & Aha, D. W. (1997). Simplifying Decision Trees:A Survey. Knowledge Engineering Review 12: 1–40.
[14] Jensen, F. (1996). An Introduction To Bayesian Networks. Springer.
[15] Gupta, B. B., & Sheng, Q. Z. (Eds.). (2019). Machine learning for computer and cyber security: principle, algorithms, and practices. CRC Press.
[16] Sahoo, S. R., & Gupta, B. B. (2020). Classification of spammer and nonspammer content in online social network using genetic algorithm-based feature selection. Enterprise Information Systems, 14(5), 710-736.
[17] Madden, M. (2003), The Performance Of Bayesian Network Classifiers Constructed Using Different Techniques, Proceedings Of European Conference On Machine Learning, Workshopon Probabilistic Graphical Models ForClassification, Pp. 59-70.
[18] Avirm Michael Kearns And Dana Ron,"Algorithmic Stability And Sanity-Check Bounds For Leave-One-Out Cross Validation".
[19] Freund, Y. (1995). Boosting A Weak Learning Algorithm By Majority. Information And Computation 121, 256285.
[20] Jiang, W. (2000). Process Consistency ForAdaboost. Tech. Report, Dept. Of Statistics,Northwestern University.
[21] Breiman, Leo, 1996. Bagging Predictors, Machine Learning.
[22] E. Alpaydin.: Voting Over Multiple Condensed Nearest Neoghbors. Artificial Intelligence Review 11:115-132, (1997) Kluwer Academic Publishers.
[23] Cristianini, N., Shawe-Taylor, 1.: An Introduction To Support Vector Machines. Cambridge University Press, Cambridge, 2000.
[24] L. Breiman, J. H. Friedman, R. A. Olshen, And C. J. Stone.Classification And Regression Trees. Wadsworth, Belmont, 1984.
[25] Sahoo, S. R., & Gupta, B. B. (2021). Real-time detection of fake account in twitter using machine-learning approach. In Advances in computational intelligence and communication technology (pp. 149-159). Springer, Singapore.