Study of Characteristics of Class of Data Mining Algorithms & Evaluate the Performance of Different Methods on Multiple Data Sets

Sai Krishna Vamshi - 2016033, Jayanth Krishna – 2016038, Sharique Ansari – 2016249

Abstract

It is generally confused among Data Analysts in selecting and applying algorithms because No Algorithm can maintain the best Performance in all data sets. Here we studied applicability of some algorithms and observed some reasons for their low accuracy in specific-type of data sets. Selecting an appropriate algorithm or exercising a class of algorithms can be done immediately with some efforts in learning the relation among attributes in data sets.

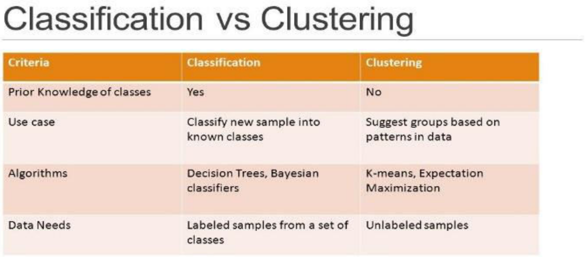
1 Introduction

Large Data collection, storage and delivery are made very easy with innovation in field of Science. To understand the value of data collected, Data Mining Algorithms are used. A wide range of Algorithms have evolved with the prosperity of Data. But their application focuses are slightly inconsistent. So, Data Analysts are required to found relatively best Data Mining method to solve problems. The comparison of varied algorithms in specific context was advised because the performance of Algorithms were proven to be problem-dependent.

2 Methods

Classification and Clustering are two types of learning techniques which portray Data Points into groups based on attributes. These two appear to be comparative yet there is some contrast between them in Data Mining context. The main difference is that Clustering is un-supervised learning where Data Points of similar features are grouped together, on the other side Classification is supervised learning technique where Data Points of Training set have pre-defined labels.

When Training is provided to the system the class label of training tuple is known before testing and then tested, this is known as supervised learning. Then again unsupervised learning does not include any training or preparation.



Classification vs Clustering

Criteria	Classification	Clustering
Prior Knowledge of classes	Yes	No
Use case	Classify new sample into known classes	Suggest groups based on patterns in data
Algorithms	Decision Trees, Bayesian classifiers	K-means, Expectation Maximization
Data Needs	Labeled samples from a set of classes	Unlabeled samples

Figure 1: Classification vs Clustering.

Classification and Clustering are the Algorithms used in Data Mining for observing Data and divide them on the basis of some classification rules or the association among Data Points. Classification classifies the data based on provided Training Data. Clustering uses different similarity measures to cluster data into groups.

2.1 Classification

Classification is a problem where a new Data Point belongs to a set of groups based on Training Set containing instances whose group it belongs to is known. An algorithm which implements classification is known as "Classifier", sometimes also refers as mathematical function that maps new Data Point to its group. Classifier performance is dependent on characteristics of the data to be classified. Different methods have been analyzed to compare classifier performance.

2.1.1 Naïve Bayes Classification

Naïve Bayes Classification is a simple probabilistic classifier by applying Bayes Theorem with assumption of independency among attributes. They use probabilities to assign the class labels. When input dimensions are high, naive Bayes classifier is used as it produce better results than many other classifiers. Let the classes be C1, C2, ...Cn. When a data point X is given, it predicts that X comes under the class which have the highest posterior probability conditioned to X. It says that X Cj belongs to Ci if and only if P(Ci/X) > P(Cj/X) where 1<=j< i. The Class Ci for which P (Ci/X) is maximized is called maximum posterior hypothesis.

P(Ci/X)= P(X/Ci)P(Ci)/P (X)

If we don't know the probabilities of the classes beforehand, we assume that the probabilities are equal. Then we maximize the P (X/Ci).

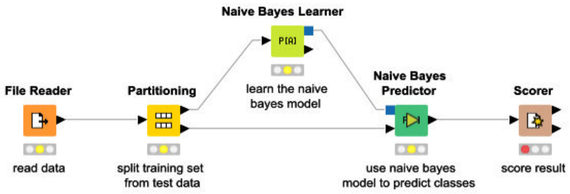

Naive Bayes Learner
P(A)
learn the naive bayes model
File Reader
read data
Partitioning
split training set from test data
Naive Bayes Predictor
use naive bayes model to predict classes
Scorer
score result

Figure 2: Naïve Bayes Classification

2.1.2 Logistic Regression

Logistic Regression is widely used statistical model for classification of binary dependent variable. Mathematically a binary logistic model has dependency variable with two possible outcomes such as win/lose, alive/dead or pass/fail. It measures the relationship between the dependent variables and one or more independent variables by estimating probabilities using a logistic method.

Generalized linear model method parameterized by θ:

$$h_\theta(X) = \frac{1}{1+e^{-\theta^T X}} = Pr(Y=1|X;\theta)$$

The basic idea behind Logistic regression is to use the mechanism developed for linear regression by modelling the Probability Pi using a linear predictor function i.e., a linear combination of the explanatory variables and a set of regression coefficients that are defined during the Training.

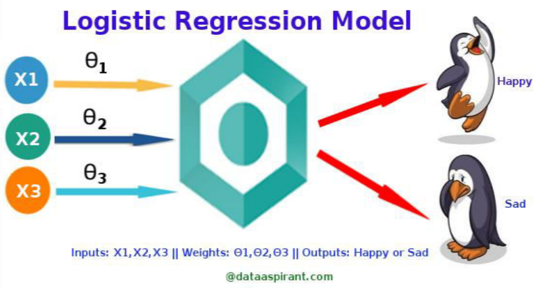

Logistic Regression Model
X1
X2
X3
θ1
θ2
θ3
Happy
Sad
Inputs: X1,X2,X3 || Weights: θ1,θ2,θ3 || Outputs: Happy or Sad
@dataaspirant.com

Figure 3: Logistic Regression

2.1.3 K-Nearest Neighbors

K – Nearest Neighbor is a non-parametric algorithm generally used for classification. In this classification the output is a class type. It is classified by a majority vote of its neighbors, with the Data Point being assigned to the type of class with most common among its K neighbors. It is a lazy learning algorithm which has no training phase.

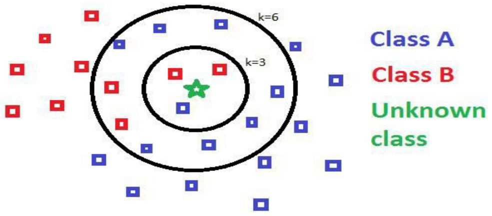


Figure 4: K-Nearest Neighbors

2.1.4 C4.5 (Statistical Classifier)

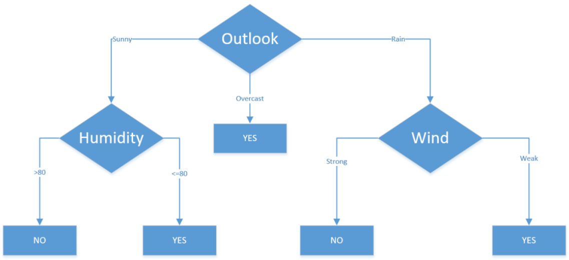
C4.5 is an extension to ID3 algorithm and solves many limitations such as numeric data, missing values faced by latter. This algorithm is used to generate Decision Trees, which are used for classification problems. It is a supervised learning algorithm, so it requires a training data set divided into attribute values and class of data point. Entropy and Information Gain are two important concepts in this algorithm and are calculated as follows

Entropy (S) = $\sum -P(I) * \log_2 P(I)$
Gain (S, A) = $entropy(s) - \sum[P(S|A) * entropy(S|A)]$

Gain ratio is calculated by dividing gain by Split Info.

Split_Info(A) = $\sum \frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|}$

Among all present attributes one with highest information gain (Gain Ratio) is chosen as parent node in tree and dataset is split according to its attributes value. This data is then passed recursively to find children nodes and complete decision tree.



C4.5 can handle both numerical and nominal values and can deal with missing values.

2.2 Clustering

Cluster analysis is the association of a gathering of examples into groups dependent on likeness or similarity. Clustering is used in various grouping, decision-making and machine learning situations. Though, in most of the conditions, some prior knowledge of data is accessible to us. So, we are allowed to make very few assumptions about the data. Clustering is utilized in a few research networks to portray strategies for gathering of unlabeled data.

2.2.1 K-Means Clustering

It is one of the primary clustering algorithms that came into existence. It works based on a very basic idea. At first, we choose k random points as our centroids and calculate the distances of every point to all the centroids and add this point to the respective centroid cluster. After one iteration, we calculate the centroids of all clusters and update the old clusters. We repeat the above step till the centroids won’t change. It can be really time consuming if we have a large data.

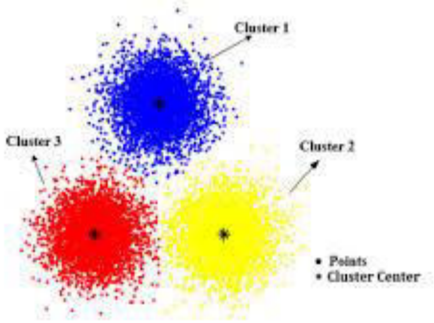


Figure 5: K-Mean Clustering

2.2.2 K-Median Clustering

It is a cluster analysis algorithm. It is a variety of k-means clustering where as opposed to figuring the mean for each cluster to decide its centroid, one rather computes the median. This has the impact of limiting mistake over all clusters regarding the 1-norm separation metric, rather than the square of the 2-norm separation metric (which k-means does.)

2.2.3 DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm. It is a density based clustering algorithm. Given an arrangement of points in some space, it groups points that are firmly packed together (points with many close-by neighbors), stamping as anomalies points that lie alone in low-density areas (whose closest neighbors are too far away)

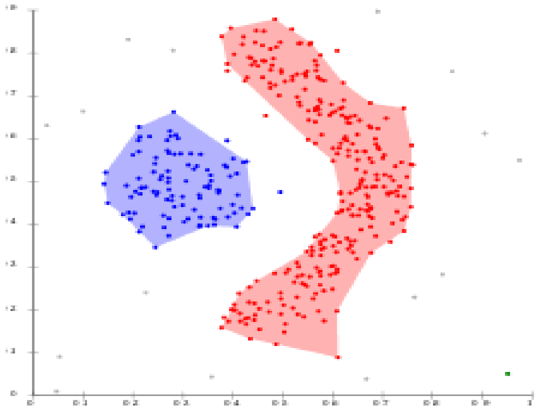


Figure 6: DBSCAN – Clusters with Different Shapes

2.3 Cluster Quality

A cluster quality measure is a function that maps pairs of the form (dataset, clustering) to some ordered set (say, the set of non-negative real numbers), so that these values reflect how 'good' or 'bad' that clustering is. Cluster quality measures may also be used to identify an ideal clustering method by comparing the different clustering solutions obtained when different clustering methods or parameters are employed over the same data set (e.g., comparing the results of a given clustering paradigm over different choices of clustering parameters, such as the number of clusters).

For the evaluation of cluster quality, we use two different measures: Entropy and Purity. These are standard measures that help us to ascertain the cluster quality.

2.3.1 Entropy

Entropy measures how the various semantic classes are distributed within each cluster. Given a particular cluster Sr of size Nr, the entropy of this cluster is defined as

$$E(S_r)=\frac{-1}{\log(q)}\sum_{i=1}^{q}\left(\frac{N_{r_i}}{N_r}\right)\left(\log\left(\frac{N_{r_i}}{N_r}\right)\right)$$

Where q is the number of classes in the dataset and Nri is the number of data points of the ith class that are assigned to the rth cluster. The entropy of the entire clustering solution is then the sum of the individual cluster entropies weighted according to the cluster size.

$$Entropy=\sum_{r=1}^{k}\left(\frac{N_r}{N}\right)E(S_r)$$

Where n is total number of data points. Small entropy value indicates better clustering solutions

2.3.2 Purity

Using the same Mathematical notation, the purity of a cluster is defined as

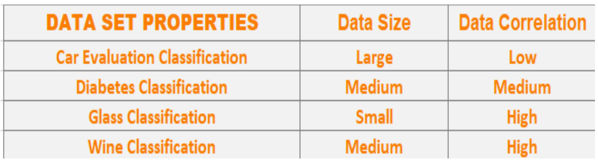
$$Pu(S_r)=\left(\frac{1}{N_r}\right)max(N_{r_i})$$

$$Purity=\sum_{r=1}^{k}\left(\frac{N_r}{N}\right)Pu(S_r)$$

Larger purity values indicate better clustering solutions. Entropy is a more comprehensive measure than purity because rather than just considering the number of documents, it considers the overall distribution of all the classes in a given cluster. For an ideal cluster with documents from only a single class, the entropy of the cluster will be 0. In general, the smaller the entropy value, the better the quality of the cluster.

3 Observations

No Algorithm can maintain the best performance in all data sets.



DATA SET PROPERTIES	Data Size	Data Correlation
Car Evaluation Classification	Large	Low
Diabetes Classification	Medium	Medium
Glass Classification	Small	High
Wine Classification	Medium	High

Figure 7: Data Set Properties

3.1 Classification

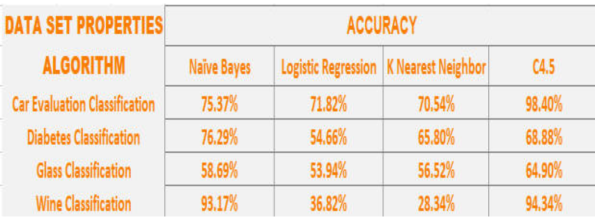
- Naïve Bayes performs very well in which Data Sets are small, less attributes, and high correlation between attributes.

- When we compare the Running time of these Algorithms Naïve Bayes is the fastest and Logistic Regression is the slowest.

- Logistic Regression is more powerful algorithm than Naïve Bayes but it requires a complex framework and it can be under fitting.

- KNN is easy to implement and it doesn't require training prior to making real time predictions. It doesn't work well with categorical features as it is not easy to calculate distance between them.

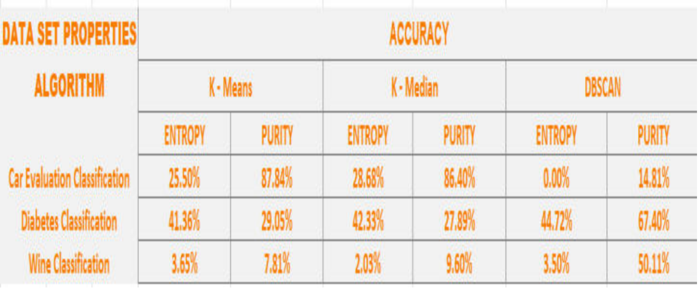
- C4.5 is a quite time efficient algorithm and handles large dataset efficiently. We can use this to solve real world problems as it can use both continuous and categorical values, and can handle missing values. But it doesn't perform very well when training datasets are small



DATA SET PROPERTIES	ACCURACY			
ALGORITHM	Naïve Bayes	Logistic Regression	K Nearest Neighbor	C4.5
Car Evaluation Classification	75.37%	71.82%	70.54%	98.40%
Diabetes Classification	76.29%	54.66%	65.80%	68.88%
Glass Classification	58.69%	53.94%	56.52%	64.90%
Wine Classification	93.17%	36.82%	28.34%	94.34%

Figure 8: Accuracy Predicted

3.2 Clustering



DATA SET PROPERTIES	ACCURACY					
ALGORITHM	K - Means		K - Median		DBSCAN	
	ENTROPY	PURITY	ENTROPY	PURITY	ENTROPY	PURITY
Car Evaluation Classification	25.50%	87.84%	28.68%	86.40%	0.00%	14.81%
Diabetes Classification	41.36%	29.05%	42.33%	27.89%	44.72%	67.40%
Wine Classification	3.65%	7.81%	2.03%	9.60%	3.50%	50.11%

Figure 8: Entropy and Purity

- Computational time of DBSCAN algorithm is less than K-Mean and K-Median.

- DBSCAN is more efficient and accurate for larger Data Sets without Noise.

- Even though we have large purity values for some clustering methods we consider Entropy as better measure for Clustering of Data Points.

References

G. Hannah Grace, Kalyani Desikan. Department of Mathematics, School of Advanced Sciences, VIT University, Chennai 600127, India. Experimental Estimation of Number of Clusters Based on Cluster Quality.

https://arxiv.org/ftp/arxiv/papers/1503/1503.03168.pdf

Zhang, Yy & Xin, Yi & Li, Qin & Ma, Jianshe & Li, Shuai & Lv, Xiaodan & Lv, Weiqi. (2017). Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. BioMedical Engineering OnLine. 16. 10.1186/s12938-017-0416-x.

https://www.researchgate.net/publication/320816509_Empirical_study_of_seven_data_mining_algorithms *_on_different_characteristics_of_datasets_for_biomedical_classification_applications*

A Data Sets Used

[0]https://www.kaggle.com/uciml/glass
[1]https://www.kaggle.com/elikplim/car-evaluation-data-set
[2]https://www.kaggle.com/rnmehta5/pima-indian-diabetes-binary-classification
[3]https://www.kaggle.com/abhikaggle8/wine-classification/data

B Image References

[0]https://i0.wp.com/sefiks.com/wp-content/uploads/2018/05/c45-result.png?w=1330&ssl=1
[1]https://i.stack.imgur.com/sxEi9.jpg
[2]https://i.stack.imgur.com/0QOII.png
[3]https://miro.medium.com/max/960/1*UgYbimgPXf6XXxMy2yqRLw.png
[4]https://bigishere.files.wordpress.com/2018/09/knn_bigisnext.jpg
[5]https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTIg7pQFJfjj8k1mbnj67VkRYDcq6x6eGwHJGd2b4S44QAI-THM