Global Journal of Computer Science and Technology
Volume 12 Issue 5 Version 1.0 March 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Content Based Data Retrieval on KNN- Classification and Cluster Analysis for Data Mining

By Aws Saad Shawkat & H K Sawant
Bharati Vidyapeeth Deemed University Pune

Abstract - Data mining is sorting through data to identify patterns and establish relationships.Data mining parameters include: Regression -In statistics, regression analysisincludes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. Sequence or path analysis -looking for patterns where one event leads to another later event. Classification -looking for new patterns. Clustering -finding and visually documenting groups. Decision Trees – Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal.

GJCST Classification: I.5.3

Strictly as per the compliance and regulations of:

Content Based Data Retrieval on KNN-Classification and Cluster Analysis for Data Mining

Aws Saad Shawkat [α] & H K Sawant [α]

Abstract - Data mining is sorting through data to identify patterns and establish relationships.Data mining parameters include: Regression - In statistics, regression analysis includes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. Sequence or path analysis - looking for patterns where one event leads to another later event. Classification - looking for new patterns. Clustering - finding and visually documenting groups. Decision Trees – Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal.

I. Introduction

Data mining is an iterative process that typically involves the following phases:

a) *Problem definition* : A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition. In the problem definition phase, data mining tools are not yet required.

b) *Data exploration* : Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital.

In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

c) *Data preparation* : Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value. In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

d) *Modeling* : Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model. In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required. The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.

e) *Evaluation* : Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions: Does the model achieve the business objective? Have all business issues been considered? At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

f) *Deployment* : Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets.

II. Data Mining Types

a) Predictive Data Mining: Predictive data mining involves creation of model system based on and described by a given set of data.

b) Descriptive Data Mining: Descriptive data mining on the other hand produces new and unique information inferred from the available set of data.

Raw Data: Raw data is a term for data collected on source which has not been subjected to processing or any other manipulation. (Primary data), it is also known as primary data. It is a relative term (see data). Raw data can be input to a computer program or used in manual analysis procedures such as gathering statistics from a survey. It can refer to the binary data on electronic storage devices such as hard disk drives (also referred to as low-level data).

Author α : Department of Information Technology Bharati Vidyapeeth Deemed University College of Engineering, Pune-46.
E-mail : awsit85@gmail.com

a) Normalization of Raw Data

Some data-mining methods, typically those that are based on distance computation between points in an n-dimensional space, may need normalized data for best results.

Here are three simple and effective normalization techniques:

b) Decimal Scaling

Decimal scaling moves the decimal point but still preserves most of the original digit value.

$$VI' = VI/10^K$$

c) Min-Max Normalization

Suppose that the data for a feature v are in a range between 150 and 250. Then, the previous method of normalization will give all normalized data between .15 and .25; but it will accumulate the values on a small subinterval of the entire range. To obtain better distribution of values on a whole, normalized interval, e.g., [0, 1], we can use the min-max formula

$$VI' = (VI - Min(VI))/(Max(VI) - Min(VI))$$

d) Standard Deviation Normalization

Normalization by standard deviation often works well with distance measures, but transforms the data into a form unrecognizable from the original data.

$$VI' = (VI - Mean(V))/Std(V)$$

Types of Data

Categorical Data: Categorical data (or variable) consists of names representing categories. For example, the gender (categories of male & female) of the people where you work or go to school; or the make of cars in the parking lot (categories of Ford, GM, Toyota, Mazda, KIA, etc) is categorical data.

Numerical Data: Numerical data (or variable) consists of numbers that represent counts or measurements. For example, the number of males & females where you work or go to school; or the number of the make of cars Ford, GM, Toyota, Mazda, KIA, etc is numerical data.

Dummy Variable: A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study.

Discrete Variable: Discrete Variable are also called Qualitative Variable. It is nominal or ordinal.

Continuous Variable: Continuous variable are measured using interval scale or ratio scale.

III. Data Reduction

Means reducing the number of cases or variables in a data matrix. The basic operations in a data-reduction process are delete column, delete a row, and reduce the number of values in a column. These operations attempt to preserve the character of the original data by deleting data that are nonessential. There are other operations that reduce dimensions, but the new data are unrecognizable when compared to the original data set, and these operations are mentioned here just briefly because they are highly application-dependent.

a) Entropy

A method for unsupervised feature selection or ranking based on entropy measure is a relatively simple technique; but with a large number of features its complexity increases significantly.

The similarity measure between two samples can be defined as

$$\alpha = -(\ln 0.5)/D$$

$$S_{ij} = e^{-\alpha D_{ij}}$$

Where Dij is the distance between the two samples xi and xj and α is a parameter mathematically expressed as

D is the average distance among samples in the data set. Hence, α is determined by the data. But, in a successfully implemented practical application, it was used a constant value of α = 0.5. Normalized Euclidean distance measure is used to calculate the distance Dij between two samples xi and xj:

$$D_{ij} = \left[\sum_{k=1}^{n}\left((x_{ik} - x_{jk})/(\max(k) - \min(k))\right)^2\right]^{1/2}$$

- where n is the number of dimensions and max(k) and min(k) are maximum and minimum values used for normalization of the k-th dimension.
- All features are not numeric. The similarity for nominal variables is measured directly using Hamming distance.

$$S_{ij} = \left(\sum_{k=1}^{n}|x_{ik} - x_{jk}|\right)/n$$

where

The total number of variables is equal to n. For mixed data, we can discretize numeric values (Binning) and transform numeric features into nominal features before we apply this similarity measure.

If the two measures are close, then the reduced set of features will satisfactorily approximate the original set. For a data set of N samples, the entropy measure is

$$E = -\sum_{i=1}^{N-1}\sum_{j=1}^{N}\left(S_{ij}\times\log S_{ij} + (1 - S_{ij})\times\log(1 - S_{ij})\right)$$

where Sij is the similarity between samples xi and xj. This measure is computed in each of the iterations as a basis for deciding the ranking of features. We rank features by gradually removing the least important feature in maintaining the order in the

configurations of data. The steps of the algorithm are base on sequential backward ranking, and they have been successfully tested on several real-world applications.

b) Linear Regreesion

In statistics, linear regression refers to any approach to modeling the relationship between one or more variables denoted y and one or more variables denoted X, such that the model depends linearly on the unknown parameters to be estimated from the data.

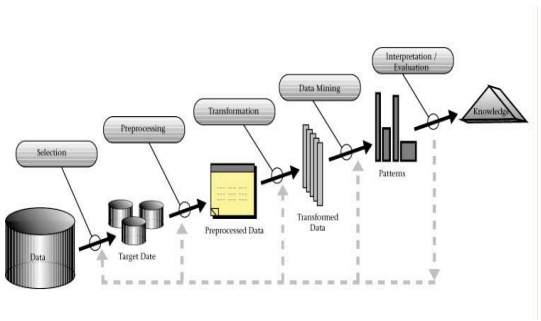
Linear regression has many practical uses. Most applications of linear regression fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, linear regression can be used to fit a predictive model to an observed data set of y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y.
- Given a variable y and a number of variables X_1, ..., X_p that may be related to y, then linear regression analysis can be applied to quantify the strength of the relationship between y and the X_j, to assess which X_j may have no relationship with y at all, and to identify which subsets of the X_j contain redundant information about y, thus once one of them is known, the others are no longer informative.

IV. Implementation

The core task of Data Mining Model is the application of the appropriate mining function to your data to build mining models that answer your business questions. Administrative tasks such as retrieving progress information or interpreting error messages support this task.

Data Mining Process



a) State the Problem

A data mining project starts with the understanding of the problem. Data mining experts and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition.

b) Data Normalization

The use of the data transformation in my project is to make the data symmetric. In practice a suitable data transformation can be selected by examining the effect of the transformation.. So for the medical data set min-max transformation is often used.

$$VI' = (VI\text{-}Min(VI))/(Max(VI)\text{-}Min(VI))$$

c) Missing Values Adjustment

The Missing value technique used in these type of project is to take the mean of that feature but the data set which I have choose for the project have no missing values.

d) Outlier Analysis

The technique used by data set to remove the outlier values is the Deviation based technique in which the human can easily distinguish unusual samples from a set of other similar samples.

After examining each and every data cluster, we obtain data set which contains no outlier.

e) Data Reduction

The term data reduction in the context o data mining is usually applied to projects where the goal is to aggregate the information contained in large data sets into manageable(smaller) information nuggets. Data reduction method can include simple tabulation ,aggregation (computing descriptive statistics) or more sophisticated technique like principle component analysis.

Since the data which I have used in the project is not so huge therefore there is no need of applying the data reduction because it could lead to the loss of information from the data.

f) Model Estimation

A model can be defined as a number of examples or a mathematical relationship. Data mining experts select and apply various mining functions because we can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types.

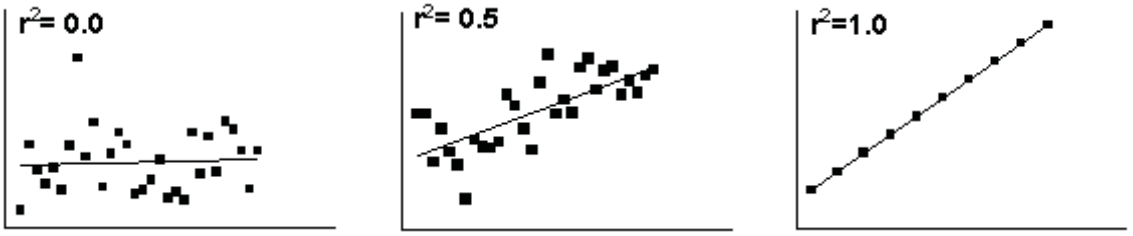
g) Linear Regression

Regression: The purpose of this model function is to map a data item to a real-valued prediction variable.

The goal of regression is to build a concise model of the distribution of the dependent attribute in terms of the predictor attributes. The resulting model is used to assign values to a database of testing records, where the values of the predictor attributes are known but the dependent attribute is to be determined.

The value r^2 is a fraction between 0.0 and 1.0, and has no units. An r^2 value of 0.0 means that knowing X does not help you predict Y. There is no linear relationship between X and Y, and the best-fit line is a horizontal line going through the mean of all Y values.

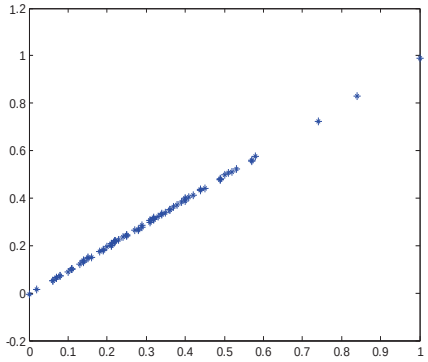
When r^2 equals 1.0, all points lie exactly on a straight line with no scatter. Knowing X lets you predict Y perfectly.



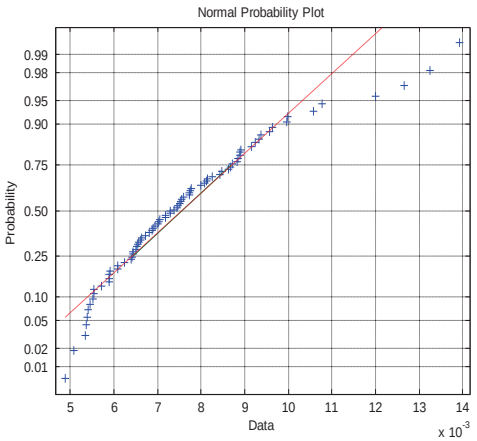
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.995[a]	.990	.989	.01950

Since the error is very small so the result which we get after applying is very close to the final result. The graph between observed and fitted value is shown in figure



The normal probability plot is a special case of the probability plot. We cover the normal probability plot separately due to its importance in many applications. The normal probability plot is formed by:
Vertical axis: Ordered response values
Horizontal axis: Normal order statistic medians
The normal probability plot is shown in the figure



h) Cluster Analysis: Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics Applying hierarchical clustering algorithm.

i) Hierarchical Clustering

It begins with as many clusters as objects. Clusters are successively merged until only one cluster remains.

- Representation of all pair-wise distances
- Parameters: none (distance measure)
- Results
- One large cluster
- Hierarchical tree (dendrogram)
- Deterministic
- Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy

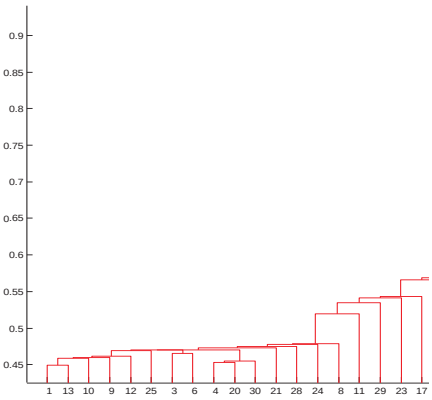
The K-means partitional-clustering algorithm is the simplest and most commonly used algorithm employing a square-error criterion.

It starts with a random, initial partition and keeps reassigning the samples to clusters, based on the similarity between samples and clusters, until a convergence criterion is met.

- Partition data into K clusters
- Parameter: Number of clusters (K) must be chosen

- Randomized initialization:
- Different clusters each time
- Non-deterministic

Here the k-mean is applied to calculate the final cluster centers among samples.



V. Conclusion

The model in which every decision is based on the comparison of two numbers within constant time is called simply a decision tree model. It was introduced to establish computational complexity of sorting and searching,__advantages of applying is Easy to understand, Map nicely to a set of business rules, Applied to real problems, Make no prior assumptions about the data, Able to process both numerical and categorical data.

Data mining techniques are used in a many research areas, including mathematics, cybernetics, genetics and marketing. Web mining, a type of data mining used in customer relationship management (CRM), takes advantage of the huge amount of information gathered by a Web site to look for patterns in user behavior.

References Références Referencias

1. Jinshu, Zhang Bofeng, Xu Xin, "Advances in Machine Learning Based Text Categorization", *Journal of Software*, Vol.17, No.9, 2006, pp.1848 1859
2. Ma Jinna, "Study on Categorization Algorithm of Chinese Text", *Dissertation of Master's Degree*, University of Shanghai for Science and Technology, 2006
3. Wang Jianhui, Wang Hongwei, Shen Zhan, Hu Yunfa, "A Simple and Efficient Algorithm to Classify a Large Scale of Texts", *Journal of Computer Research and Development*, Vol.42, No.1, 2005, pp.85 93
4. Li Ying, Zhang Xiaohui, Wang Huayong, Chang Guiran, "Vector-Combination-Applied KNN Method for Chinese Text Categorization", *Mini- Micro Systems*, Vol.25, No.6, 2004, pp.993 996
5. Wang Yi, Bai Shi, Wang Zhang'ou, "A Fast KNN Algorithm Applied to Web Text Categorization", *Journal of The China Society for Scientific and Technical Information*, Vol.26, No.1, 2007, pp.60 64
6. Fabrizio Sebastiani, "Machine learning in automated text categorization", *ACM Computer Survey*,Vol.34, No.1, 2002, pp. 1-47
7. Lu Yuchang, Lu Mingyu, Li Fan, "Analysis and construction of word weighing function in vsm", *Journal of Computer Research and Development*, Vol.39, No.10, 2002, pp.1205 1210
8. Belur V, Dasarathy, "Nearest Neighbor (NN) Norms NN Pattern Classification Techniques", *Mc Graw-Hill Computer Science Series, IEEE Computer Society Press*, Las Alamitos, California, 1991,pp.217-224
9. Yang Lihua , Dai Qi, Guo Yanjun, "Study on KNN Text Categorization Algorithm", *Micro Computer Information*, No.21, 2006, pp.269 271
10. Wang Yu, Wang Zhengguo, „A fast knn algorithm for text categorization", *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics*, Hong Kong, 19-22 August 2007, pp.3436-3441
11. Yang Y, Pedersen J O, "A comparative study on feature selection in text categorization", *ICNL*, 1997,pp.412-420
12. Xinhao Wang, Dingsheng Luo, Xihong Wu, Huisheng Chi, "Improving Chinese Text Categorization by Outlier Learning", *Proceeding ofNLP-KE'05* pp. 602-607
13. Jin Yang, Zuo Wanli, "A Clustering Algorithm Using Dynamic Nearest Neighbors Selection Model", *Chinese Journal of Computers*, Vol.30, No.5, 2005, pp.759 762

This page is intentionally left blank