ISSN 2277-3061

THE IMPORTANCE OF NORMALIZATION METHODS FOR MINING MEDICAL DATA

Mihaela Gheorghe and Ruxandra Petre
Faculty of Economic Cybernetics, Statistics and Informatics, Bucharest University of Economic Studies, Romania
Mihaela.Gheorghe@ie.ase.ro
Faculty of Economic Cybernetics, Statistics and Informatics, Bucharest University of Economic Studies, Romania
Ruxandra_Stefania.Petre@yahoo.com

ABSTRACT

Over the past decades, the field of medical informatics has been growing rapidly and has drawn the attention of many researchers. The digitization of different medical information, including medical history records, research papers, medical images, laboratory analysis and reports, has generated large amounts of data that need to be handled. As the rate of data acquisition is greater than the rate of data interpretation, new computational technologies are needed in order to manage the resulted repositories of medical data and to extract relevant knowledge from them. Such methods are provided by data mining techniques, which are used for discovering meaningful patterns and trends within the data and help improving various aspects of health informatics. In order to apply data mining techniques, the data needs to be cleansed and transformed, normalization being one of the most important pre-processing methods that accomplish this purpose.

This paper aims to present the impact of applying different data normalization methods, on the performance obtained with the K-Nearest Neighbour algorithm on medical data sets.

Keywords

Medical informatics, Data mining, Pre-processing, Normalization, K-Nearest Neighbour

Academic Discipline And Sub-Disciplines

Computer Science, Data Mining, Medical Informatics

SUBJECT CLASSIFICATION

Machine learning

TYPE (METHOD/APPROACH)

Experimental

Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol.14, No.8
www.ijctonline.com, editorijctonline@gmail.com

ISSN 2277-3061

INTRODUCTION

Nowadays, the amount of data that is collected and needs to be processed is growing exponentially every day. In the medical field, different researchers and companies are mining huge amounts of data in order to get to proper conclusions for their case studies.

Physicians and other medicine practitioners are providing treatment recommendations to patients based on their medical history, laboratory results, medical images etc. Health informatics can help them by providing a more rapid access to the relevant information that need to be reviewed and thus, allowing them to make optimal decisions. Moreover, by integrating data mining techniques in the health informatics solutions, physicians will be able to use analytical and predictive instruments to query all the relevant information that are useful within their decision process.

Data mining techniques are used for their ability to provide the necessary methods and tools for extracting meaningful patterns and trends, based on the data gathered within the databases taken into consideration. In order for data mining algorithms to be applied, data needs to be validated, cleaned and transformed. This is achieved as part of a preliminary step of the data mining process – pre-processing. One of the core methods used during pre-processing is represented by data normalization.

In the process of developing informatics solutions for the medicine field, it is very important to pre-process all the required medical data sets before applying any data mining algorithms. This is due to the fact that data is mostly characterized by noise, discrepancies, outliers, missing values and lack of exactness. In order to run a successful analysis on medical data sets, such as K-NN algorithm, different pre-processing methods should be carried out first.

K-Nearest Neighbour (k-NN) is a classification algorithm that uses mostly the concept of Euclidian distance, although other distance measures can be used as well, in order to classify an input data based on the class label of the closest k points in the training dataset [1]. The concept of closeness between points is influenced by the dataset accuracy as well. Thus, the normalization method needs to be applied in order to reduce the redundant data, for transforming the initial data set into a more consistent and noise-free one and, overall, to ensure that good quality clusters are generated at the end of the performed analysis.

NORMALIZATION: THEORETICAL FRAMEWORK

Data mining is the core step of the Knowledge Discovery in Databases (KDD) process. In order to apply data mining techniques to the data, a crucial issue, part of the KDD process, must be addressed – data pre-processing. The KDD process flow, along with its main components shown in Figure-1:

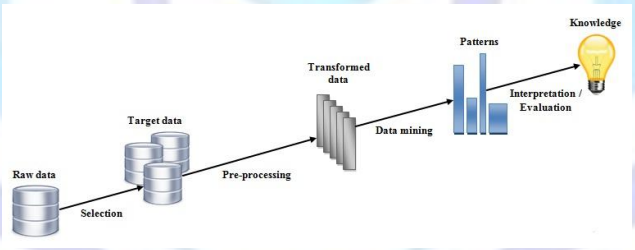

Fig 1: KDD process flow

According to the KDD process flow, after the raw data is selected from the sources and loaded as target data into the mining database, it has to be cleansed and transformed, so that data mining techniques can be applied to it, to obtain significant patterns and trends. These patterns and trends are evaluated and they become valuable knowledge extracted from the data.

Data pre-processing involves various methods divided into four major categories: [2]

- **data cleaning** – attempts to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data;
- **data integration** – implies merging the data from multiple data stores, to reduce and avoid redundancies and inconsistencies;
- **data reduction** – is applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data;
- **data transformation** – the data are transformed or consolidated into forms appropriate for applying data mining techniques.

The main data transformation methods used in data mining are: smoothing, attribute construction, normalization, aggregation and discretization.

ISSN 2277-3061

The major data pre-processing categories, along with the methods used for data transformation and the main data normalization methods are shown in Figure-2:

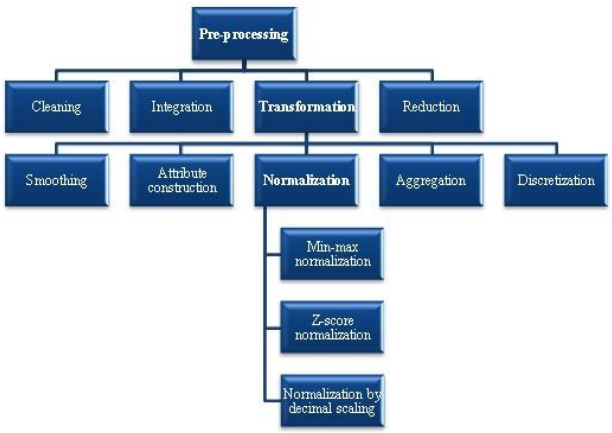

Fig 2: Data pre-processing methods

Data normalization method is used to reduce the range of an attribute of the dataset to a smaller range, for example 0 to 1.0.

Normalization is used to standardize all the features of the dataset into a specified predefined criterion so that redundant or noisy objects can be eliminated and use made of valid and reliable data which can effect and improve accuracy of the result [3].

There are several data normalization methods, the most important ones being min-max normalization, z-score normalization and normalization by decimal scaling: [2]

1. **Min-max normalization** performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v_i, of A to v_i' in the range [new_minA, new_maxA] by computing (1).

$$v_i' = \frac{v_i - min_A}{max_A - min_A}\left(new\_max_A - new\_min_A\right) + new\_min_A \quad (1)$$

where,
- v_i' = new value of attribute A
- v_i = current value of attribute A
- min_A = minimum value of attribute A
- max_A = maximum value of attribute A
- new_min_A = minimum value for new range of attribute A
- new_max_A = maximum value for new range of attribute A

Min-max normalization preserves the relationships among the original data values. It will encounter an "out-of-bounds" error if a future input case for normalization falls outside of the original data range for A.

2. **Z-score normalization** (or zero-mean normalization), the values for an attribute, A, are normalized based on the mean (i.e. average) and standard deviation of A. A value, v_i, of A is normalized to v_i' by computing (2).

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A} \quad (2)$$

where,

- v_i' = new value of attribute A
- v_i = current value of attribute A
- $\bar{A}$ = mean of attribute A
- σ_A = standard deviation of attribute A

This method of normalization is useful when the actual minimum and maximum of attribute A are unknown, or when there are outliers that dominate the min-max normalization.

3. **Normalization by decimal scaling** normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value, v_i, of A is normalized to v_i' by computing (3).

$$vi' = \frac{vi}{10^{j}} \qquad (3)$$

where,

- v_i' = new value of attribute A
- v_i = current value of attribute A
- j = the smallest integer such that max($|v_i'|$) < 1

Data normalization is very useful in data mining for classification algorithms, involving neural networks or distance measurements such as nearest-neighbor classification – k-NN.

CASE STUDY

In this chapter we will describe the results obtained after applying the k-NN algorithm's steps on each dataset that resulted after cleaning the initial Diabetes dataset by one of the normalization methods.

The dataset used for this case study is represented by the instances associated to patients investigated for Diabetes. Those were collected from the National Institute of Diabetes and Digestive and Kidney Diseases, from India. The dataset was obtained from UCI Machine Learning Repository [4] and consists in a number of 768 instances. In what concerns the structure of the dataset, this is represented by 8 attributes and a class, illustrated in Table-1. The class labels can take two values: "tested positive" and "tested negative" and represents the diagnosis set for each patient investigated as having diabetes.

Table 1. Diabetes dataset structure

Attribute name	Unit
Number of times pregnant	Numeric
Plasma glucose concentration	mg/dl
Triceps skin fold thickness	mm
Diastolic blood pressure	mmHg
2-Hour serum insulin	μU/ml
Body mass index	((weight in kg/(height in m)^2)
Diabetes pedigree function	Numeric
Age	Numeric (years)
Class variable	String (tested positive or tested negative)

During our experimental study, the k-NN algorithm was applied, using RapidMiner [5], on the dataset described above, normalized using the methods previously presented.

RapidMiner is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics.

In terms of data normalization, RapidMiner Software documentation defines this preprocessing technique as a tool used to rescale attribute values to fit in a specific range. Normalization of the data is very important when dealing with attributes of different units and scales, especially for data mining techniques that use the Euclidean distance. Therefore, all attributes should have the same scale for a fair comparison between them. In other words normalization is a technique used to level the playing field when looking at attributes that widely vary in size as a result of the units selected for representation. The RapidMiner normalization operators perform normalization of selected attributes. Four normalization methods are provided by RapidMiner: range_transformation, proportion_transformation, z_transformation and interquartile_range. [5]

Our experiment consisted in defining a mining process in RapidMiner, for applying the k-NN algorithm on the Diabetes dataset, after applying each normalization method, and obtaining performance indicators for each execution of the process.

The focus of this process is to obtain a comparative analysis of different methods available for normalization. All normalization parameters other than the method parameter are for selection of attributes on which normalization is to be applied.

The process flow for applying the k-NN classification algorithm is shown in Figure-3:

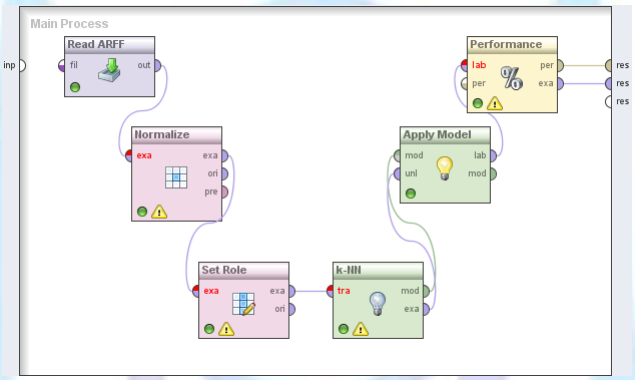

Fig 3: RapidMiner process flow

The process we defined for testing the performance of k-NN, has the following six steps:

1. Read ARFF – this operator is used for reading an ARFF file, in our case the file diabetes.arff, for processing purposes.
2. Normalize – this operator normalizes the attribute values of the selected attributes of the dataset. RapidMiner supports applying all three normalization methods described in this paper. We applied during our experiment the following RapidMiner methods: for min-max normalization we used range_transformation, for z-score normalization we used z_transformation and for normalization by decimal scaling we used proportion_transformation. All normalization parameters other than the method parameter are for selection of attributes on which normalization is to be applied.
3. Set Role – this operator is used to change the role of one or more attributes. The Role of an attribute reflects the part played by that attribute in the dataset. Changing the role of an attribute may change the part played by that attribute in a process.
4. k-NN – this operator generates a k-Nearest Neighbour model from the input dataset. This model can be a classification or regression model depending on the input dataset. The basic k-NN algorithm is composed of two steps: find the k training examples that are closest to the unseen example and take the most commonly occurring classification for these k examples.
5. Apply Model – this operator applies an already learnt or trained model, in our case k-NN, on a dataset. A model is first trained on an dataset; information related to the dataset is learnt by the model. Then that model can be applied on another dataset usually for prediction.
6. Performance – this operator is used for performance evaluation. It delivers a list of performance criteria values. These performance criteria are automatically determined in order to fit the learning task type, for our experiment k-NN.

The process described above was executed for each normalization method on the Diabetes dataset. The results obtained have shown the performance of the k-NN algorithm on the dataset, transformed through each normalization method.

The effect of training the k-NN algorithm on the described database can be measured with a set of statistical indicators. For this study, we have taken into consideration the accuracy and root mean square error which are representative in terms of describing an algorithm's efficiency and in order to evaluate and compare the impact that each normalization methods has on the results. Equations (4) and (5) are used to calculate the root mean square error (RMSE) and accuracy (Acc):

1. **Root mean square error**

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{ci} - x_{ri})^2}{n}} \qquad (4)$$

where,

- x_{ci} = the calculated value based on the algorithm
- X_{ri} = the real value from the database
- n = the number of instances used for training and testing the k-NN algorithm

2. **Accuracy**

$$AC = \frac{a_{11} + a_{22}}{a_{11} + a_{12} + a_{21} + a_{22}} \qquad (5)$$

where,

- a_{11} = the number of negative instances correctly identified as negative
- a_{12} = the number of negative instances incorrectly identified as positive
- a_{21} = the number of positive instances incorrectly identified as negative
- a_{22} = the number of positive instances correctly identified as positive

Data transformation methods such as normalization are meant to increase the efficiency and accuracy of data mining techniques, such as k-NN algorithm. As part of our experiment we applied all three normalization methods and we performed a comparative analysis against each other. The purpose of the experiment was to see the effect of the three normalization techniques applied over the accuracy and root mean square error of k-NN.

In order to establish the better normalization technique based on our experiments, the root mean square error metric should be as low as possible and in the same time, the accuracy of the classification process should be higher. These expectations are achieved with the min-max normalization method as shown in Table-2:

Table 2. Comparative analysis of the normalization methods

Normalization method	Accuracy	Root mean squared error
Min-max normalization	85.81%	0.312
Z-score normalization	85.16%	0.318
Normalization by decimal scaling	82.94%	0.339

From the table above, it is clear that min-max normalization produces the lowest RMSE compared to the other methods. It is followed by Z-score normalization where the RMSE is 0.318. The highest RMSE is obtained through normalization by decimal scaling.

In regards to the prediction accuracy, with min-max normalization it is 85.81%, followed by Z-score normalization with 85.16% and normalization by decimal scaling with 82.94%.

Our experiment compares the quality of the classification obtained by applying k-NN algorithm, on the diabetes dataset, with min-max normalization method against the two other normalization methods, Z-score normalization and normalization by decimal scaling. Based on the previous assertions, all three normalization procedures produce almost

same quality groups for Diabetes dataset, but both min-max normalization and Z-score normalization methods yield better performance than Decimal Scaling for this dataset.

CONCLUSIONS

Within this paper we have developed a systematic study regarding the influence of each normalization method on the results obtained after running a classification algorithm in order to train and validate the 'Diabetes' medical dataset. We can conclude, based on the presented results, that an adequate normalization of input data prior to training and testing k-NN algorithm generates two main advantages. This approach allows us to estimate and classify new instances with a better accuracy and also, the estimation errors that represent efficiency metric can be reduced. Therefore, an adequate normalization method is represented by a linear scale transformation that changes the same attributes values to the same relative variation after applying a min-max domain interval conversion.

REFERENCES

[1] Amit Dhurandhar and Alin Dobra, "Probabilistic Characterization of Nearest Neighbor Classifier", Available online (May 2nd, 2015): http://researcher.watson.ibm.com/researcher/files/us-adhuran/nnjMLC.pdf

[2] J. Han, Micheline Kamber and Jian Pei, "Data Mining: Concepts and Techniques. Third Edition", Morgan Kaufmann Publishers, USA, 2011, ISBN 978-0-12-381479-1.

[3] Vaishali R. Patel and Rupa G. Mehta, "Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2, September 2011, pp. 331-336, ISSN (Online): 1694-0814, Available online (May 3rd, 2015): http://ijcsi.org/papers/IJCSI-8-5-2-331-336.pdf

[4] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. Available online (May 6th, 2015): https://archive.ics.uci.edu/ml/datasets/Diabetes

[5] RapidMiner Software. Available online (May 6th, 2015): https://rapidminer.com/

Authors

Mihaela GHEORGHE has graduated the Faculty of Economic Cybernetics, Statistics and Informatics of the Bucharest University of Economic Studies in 2010 as a promotion leader. In 2012 she graduated the Informatics Economics Master program also as a promotion leader. She is currently conducting research in Economic Informatics at Bucharest University of Economic Studies, coordinated by Professor dr. Bogdan Ghilic-Micu for her PhD thesis "Architecture and technologies in telemedicine". She was a pre-Assistant within the Department of Economic Informatics between 2012 and 2013. Currently she is working as a software engineer at IBM Corporation, since 2012. Her main scientific interests include: telemedicine systems, mobile technologies, programming and artificial neural networks.

Ruxandra PETRE graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies in 2010. In 2012 she graduated the Business Support Databases Master program. Currently, she is a PhD candidate, coordinated by Professor Ion LUNGU in the field of Economic Informatics at the Bucharest University of Economic Studies. Her scientific fields of interest include: Databases, Data Warehouses, Business Intelligence, Decision Support Systems and Data Mining.