Open Comput. Sci. 2017; 7:36–40 DE GRUYTER OPEN

Research Article

Joana Pereira, Hugo Peixoto*, José Machado, and António Abelha

A Data Mining Approach for Cardiovascular Diagnosis

https://doi.org/10.1515/comp-2017-0007
Received Nov 17, 2017; accepted Nov 29, 2017

Abstract: The large amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analysed by traditional methods. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data. In the healthcare industry specifically, data mining can be used to decrease costs by increasing efficiency, improve patient quality of life, and perhaps most importantly, save the lives of more patients. The main goal of this project is to apply data mining techniques in order to make possible the prediction of the degree of disability that patients will present when they leave hospitalization. The clinical data that will compose the data set was obtained from one single hospital and contains information about patients who were hospitalized in Cardio Vascular Disease's (CVD) unit in 2016 for having suffered a cardiovascular accident. To develop this project, it will be used the Waikato Environment for Knowledge Analysis (WEKA) machine learning Workbench since this one allows users to quickly try out and compare different machine learning methods on new data sets

Keywords: healthcare information systems; knowledge discovery; data mining; machine learning; classification algorithms; cerebrovascular accidents

1 Background

Data mining has been used in many industries to improve customer experience and satisfaction, and increase product safety and usability. In healthcare, data mining has proven effective in areas such as predictive medicine, customer relationship management, detection of fraud and abuse, management of healthcare and measuring the effectiveness of certain treatments. For these reasons, data mining is becoming increasingly popular and essential in this area [1]. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data. In addition, several other factors have motivated the use of data mining applications in healthcare [2].

The process of applying computer based information system (CBIS), including new techniques, for discovering knowledge from data is called data mining [3, 4]. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. Thus, data mining can be defined as being a process of non-trivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database [5]. This process uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used [6].

Data mining tasks can be divided into descriptive and predictive. While descriptive tasks aim to find a human interpretation of forms and associations, after reviewing the data and the entire construction of the model, prediction tasks tend to predict an outcome of interest. Classification and regression are examples of predictive tasks while clustering and association are descriptive tasks [7–9].

It is important to discuss the relationship between data mining and Knowledge Discovery in Databases (KDD) due to their similarity in processes and outcomes [10, 11]. Knowledge Discovery (KDD) is a process that allows automatic scanning of high-volume data in order to find useful patterns that can be considered knowledge about the data. Thus, although data mining and KDD are often treated as equivalent, actually, data mining is an important step in the KDD process [12, 13]. The additional steps in the KDD process, such as data selection, data pre-processing, data transformation, and proper interpretation/evaluation of the results of mining, ensure that useful knowledge is derived from the data. On the other hand, data mining is

*Corresponding Author: Hugo Peixoto: Algoritmi Research Center, University of Minho, Campus Gualtar, Braga 4710, Portugal; Email: hpeixoto@di.uminho.pt
Joana Pereira: University of Minho, Campus Gualtar, Braga 4710, Portugal
José Machado, António Abelha: Algoritmi Research Center, University of Minho, Campus Gualtar, Braga 4710, Portugal

the step that allows the extraction of patterns from preprocessed data through the applications of specific algorithms [10, 11, 14].

The main goal of this project is to apply data mining techniques in order to make possible the prediction of the degree of disability that patients will present when they leave hospitalization. To develop this project, it will be used the WEKA machine learning Workbench since this one allows users to quickly try out and compare different machine learning methods on new data sets. Its modular, extensible architecture allows sophisticated data mining processes to be built up from the wide collection of base learning algorithms and tools provided [15]. WEKA can process data given in the form of a single relational table. Its main objectives are to assist users in extracting useful information from data and enable them to easily identify a suitable algorithm for generating an accurate predictive model from it [16, 17].

2 Materials and Methods

In this project, a previous existent Data Warehouse was used to construct the data set. It is important to note that the clinical data that compose the Data Warehouse was obtained from one single hospital and contains information about patients who were hospitalized in CVD's unit in 2016 for having suffered a cardiovascular accident.

After extracting the data to Microsoft Excel it was necessary to normalize it. It is important to note that, before proceeding to the data normalization, it was necessary to remove all the rows that had columns with empty or unknown values. After deleting these rows, the data were reduced from 447 patients to 178.

Then, the file was saved in Comma Separated Value (.csv) format in order to load data into WEKA, it's necessary to put it into a format that it can interpret. Although WEKA accepts files in CSV format, its preferred method for loading data is in the Attribute-Relation File Format (.arff), where you can define the type of data being loaded then supply the data itself. This format is an extension of the CSV file format where a header is used that provides metadata about the data types in the columns. So, it was used a handy tool provide by WEKA to load the CSV file and save it in ARFF.

The first data set loaded to WEKA (data set 1) consisted of data from 178 patients and contained information about the patient's sex, age, previous ranking, clinical classification, risk factors, type of stroke and exit ranking. In addition, information about three periods of time were also added: the time that elapsed from the moment the patient felt the first symptoms until the moment he was admitted in the hospital – *time_symptom_door*; the time between the patient's entry on the hospital and his, or her, admission into the neurology service – *time_door_neurology*; and, finally, the time between the patient's arrival at the hospital and the time at which the CT scan was performed – *time_door_CT*. It has to be noted that a CT scan must be performed on all patients in order to find out the type of stroke they suffered. After loading the ".arff" file to WEKA, the "NumericToNominal" filter was applied to enforce the Exit Ranking attribute to become nominal.

Finally, different algorithms were applied to the data set in order to perceive which was the most appropriate and that allowed the obtainment of better results. The algorithms tested were WEKA's classification algorithms since the main objective of this project is related to the prediction of the patient's exit ranking. Exit ranking values are predefined nominal values that express the degree of disability patients present when they leave hospitalization and are divided into seven possible classifications: asymptomatic, non-disabling symptoms, light disability, moderate disability, moderately severe disability, severe disability and death. In fact, after eliminating the rows that had empty or unknown values, the data referring to patients associated to the exit ranking "non-disabling symptoms" ceased to exist. This is the reason why, at the moment of the confusion matrix visualization, only six classifications are recognized.

As mentioned before, several algorithms were tested during the development of this project. In order to perceive those who presented better performance, two factors were taken into account: the accuracy (percentage of correctly classified instances) and the relative absolute error (mean absolute error divided by the error of the *ZeroR* classifier).

It is also important to mention the operation of the *ZeroR* algorithm since this is the most primitive learning algorithm in WEKA. It models the dataset with a single rule. Given a new data item for classification, *ZeroR* always predicts the most frequent category value in the training data for problems with a nominal class value, or the average class value for numeric prediction problems.

Although it seems to make little sense to use this algorithm for classification, it can be useful for generating a baseline performance that other learning schemes are compared to. In some datasets, it is possible for other learning schemes to induce models that perform worse on new data than *ZeroR* which is a clear indicator of serious overfitting. It should be noted that the accuracy value obtained after applying this algorithm to the data set was

25.8427% and that the absolute relative error value was, as expected, 100%.

3 Results

After applying several classification algorithms to the previously mentioned data set it was possible to observe that the percentage of correctly classified instances, as well as the confusion matrix, were not favorable. In fact, the best result was obtained through the *lazy.KStar* algorithm and is associated with an accuracy of 73.0337%.

It is important to note that, in addition to the percentage of correctly classified instances being extremely low, the relative absolute error associated with this result is 34.847%. Thus, in order to improve these results, several experiments were performed, eliminating different attributes in the sense of perceiving if there were significant changes that could enhance the final results.

During the elimination of the attributes it was possible to verify that the type of stroke that the patient suffered has no influence on its exit ranking prediction. The same is true in what concerns the attributes *time_door_neurology* and *time_door_CT* and some of the attributes related to the patient's risk factors. On the other hand, the elimination of any of the other attributes will worsen the results. It is important to refer that the attribute that clearly has more influence on the results is the one related to the patient's clinical classification.

Then, since the elimination of different attributes did not allow the improvement of results, it was decided to change the data set. This change consisted in decreasing the number of exit ranking classifications, causing an increase in the number of patients associated to each of them.

Therefore, the possible classifications were reduced to five: without disability, light disability, moderate disability, severe disability, and death (data set 2). These changes had the desired effect since both the accuracy and the confusion matrix improved significantly. The algorithm that allowed the obtainment of better results was the *tree.RandomTree* algorithm. Applying this algorithm, percentages of correctly classified instances and relative absolute error of 78.6517% and 29.2593%, respectively, were obtained.

However, although there is a significant improvement of the results, they continue to present relatively low percentages of correctly classified instances and very high values of relative absolute error. Therefore, since the approach to reduce the number of classifications associated with the patient's exit ranking was effective, the referred classifications where, this time, limited to: without disability, with disability and death. It should be noted that the classifications "without disability" and "death" remained exactly the same and that the only change was to combine previous "light disability", "moderate disability" and "severe disability" classifications into a single classification - "with disability" (data set 3).

As expected, after this change, the percentage of correctly classified instances increased significantly. However, the relative absolute error also increased. The best result was obtained by applying the *meta.RandomCommittee* algorithm. Through this algorithm, a percentage of correctly classified instances of 89.8876% and a relative absolute error of 35.7958% were obtained.

In order to improve the results, it was decided to duplicate the cases associated with the classifications "without disability" and "death" – this technique is called oversampling (data set 4). After uploading the new data set into WEKA, it was verified, as expected, an increase of the percentage of correctly classified instances up to 93.5185% and a significant decrease of the relative absolute error - which took the value of 12.4723%. The algorithm that allowed to obtain these results was the *trees.RandomTree*.

The *meta.RandomCommittee* algorithm allowed to obtain equally favorable results. The percentages of correctly classified instances and relative absolute error obtained through this algorithm were 94.9074% and 14.2541%, respectively.

Finally, a last data set was tested since the previous ones involved the modification of the original data. Thus, in order to keep the exit ranking classifications defined in the initial data, the data that constituted the original data set was duplicated (data set 5). After applying the first classification algorithm to the new data set, consisting of original duplicated data, it was immediately possible to verify a significant increase in the quality of the results obtained. The *trees.LMT* algorithm even allowed an accuracy of 98.3146% and a relative absolute error of 5.1039%. Equally good results were obtained using the *trees.RandomTree* algorithm. The percentages of correctly classified instances and relative absolute error obtained through this algorithm were 97.7528% and 2.7486%, respectively. Table 1 summarizes the best result obtained for each one of the above-mentioned data sets.

Table 1: Best performance algorithms by accuracy

Algorithm	Accuracy	Relative Absolute Error	Data Set
lazy.KStar	73.0337%	34.847%	1
trees.RandomTree	78.651%	29.2593%	2
meta.RandomCommittee	89.8876%	35.7958%	3
trees.RandomTree	93.5185%	12.4723%	4
meta.RandomCommittee	94.9074%	14.2541%	
trees.LMT	98.3146%	5.1039%	5
trees.RandomTree	97.7528%	2.7486%	

4 Discussion and Conclusions

As mentioned above, the results obtained by applying algorithms to the original data set weren't favorable. In this sense, several changes were performed to the previously mentioned data set. The first approach consisted in eliminating different attributes in the sense of perceiving if there were significant changes that could enhance the final results. This process allowed to verify that, contrary to what would be expected, the time that elapsed from the moment that the patient arrived at the hospital and the time at which the CT scan was performed has no influence on the patient's exit ranking prediction.

The same results were obtained for the attributes related to the type of stroke that the patient suffered and to the time that elapsed from the moment that the patient arrived at the hospital and the moment of his or her admission into the neurology service. These results were equally unexpected, since it would be expected that the type of stroke had influence on the patient's exit ranking. In addition, the patient's entry into the neurology service corresponds to the moment in which it is perceived that the patient suffered a stroke. In this sense, it would be normal that the faster the patient enters in the neurology service, the faster the CT scan would be performed and, for the reasons mentioned above, the more favorable would be his or her exit ranking.

Since the elimination of different attributes did not allow the improvement of results, the initial data set were changed. Once the approach of reducing the number of classifications associated with the patient's exit ranking was effective, the next step was to limit the referred classifications to three. As expected, after this change, the percentage of correctly classified instances increased significantly. However, the relative absolute error has also increased. This is because the number of cases associated with "without disability" and "death" classifications is very small compared to the number associated with the classification "with disability".

Most likely, this makes the vast majority of training cases fall on the data of patients with an exit ranking of "with disability". Therefore, the algorithm errs slightly more than previously when classifying the data associated with the patient's exit rankings "without disability" and "death". This problem could be resolved if there were more cases associated with these two classifications. Thus, even if it isn't "ideal", it was decided to duplicate the cases associated with the mentioned classifications – this technique is called oversampling. As expected, after this change, the results improved significantly.

As discussed, in order to obtain favorable results, it was necessary to change the original data set. This procedure could probably have been avoided if the amount of data constituting the original data set was greater since it consisted of data of only 178 patients. That is the reason why a last data set was tested. In order to keep the exit ranking classifications defined in the initial data, the data that constituted the original data set was duplicated. This data set was the one that allowed the obtainment of the most favorable results.

In conclusion, after analyzing the results, it is possible to conclude that the algorithm that allowed the obtainment of better results was the *trees.RandomTree* algorithm. Although this algorithm is not always the one associated with the best values of accuracy, it is the one that allowed to obtain the smallest values of relative absolute error and, at the same time, it still presents high percentages of correctly classified instances.

It should be noted that a higher relative absolute error is associated with a classification more distanced from the reality. Thus, in the context of this project, the trees.RandomTree algorithm being associated with the lower values of relative absolute error is extremely important since we are dealing with a context that is directly associated, as explained in the previous paragraph with the patient's lives.

Some difficulties were encountered during the development of this project since, in order to obtain favorable results, it was necessary to modify the original data set. These changes consisted, essentially, in decreasing the number of exit ranking classifications, causing an increase in the number of patients associated to each of them; and, in duplicating the data associated with the classifications "without disability" and "death". This procedure could probably have been avoided if the amount of data constituting the data set was greater. It should be noted that the data set used consists of data of only 178 patients.

Acknowledgement: This work has been supported by Compete: POCI-01-0145-FEDER-007043 and FCT within the Project Scope UID/CEC/00319/2013.

References

[1] USF Health, Data Mining In Healthcare. https://www.usfhealthonline.com/re-sources/healthcare/data-mining-in-healthcare/ (Online). (Accessed in: 26-05-2017).

[2] Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. Journal of healthcare information management, 19(2), 65.

[3] Jothi, N., & Husain, W. (2015). Data mining in healthcare – a review. Procedia Computer Science, 72, 306-313.

[4] Vlahos, G. E., Ferratt, T. W., & Knoepfle, G. (2004). The use of computer-based information systems by German managers to support decision making. Information & Management, 41(6), 763-779.

[5] Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2013). A Study On Clinical Prediction Using Data Mining Techniques. International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), 1(3), 239-248.

[6] Durairaj, M., & Ranjani, V. (2013). Data mining applications in healthcare sector: a study. Int. J. Sci. Technol. Res. IJSTR, 2(10).

[7] Kharya, S. (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. arXiv preprint arXiv:1205.1923

[8] Kohavi, R., & Quinlan, J. R. (2002, January). Data mining tasks and methods: Classification: decision-tree discovery. In Handbook of data mining and knowledge discovery (pp. 267-276). Oxford University Press, Inc.

[9] Wide Skills, Data Mining Tasks. http://www.wideskills.com/data-mining-tutorial/05-data-mining-tasks (Online). (Accessed in: 28-05-2017).

[10] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems, 36(4), 2431-2448.

[11] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

[12] Goebel, M., & Gruenwald, L. (1999). A survey of data mining and knowledge discovery software tools. ACM SIGKDD explorations newsletter, 1(1), 20-33.

[13] ȚĂRANU, I. (2016). Data mining in healthcare: decision making and precision. Database Systems Journal BOARD, 33.

[14] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

[15] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

[16] Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. Bioinformatics, 20(15), 2479-2481.

[17] pentaho, A Hitachi Group Company, Data Mining – Weka. http://community.pen-taho.com/projects/data-mining/ (Online). (Accessed in: 26-05-2017).