MultiCraft

International Journal of Engineering, Science and Technology
Vol. 12, No. 3, 2020, pp. 81-89

INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE AND TECHNOLOGY

www.ijest1-ng.com
www.ajol.info/index.php/ijest

Comparison of the accuracy of classification algorithms on three data-sets in data mining: Example of 20 classes

T. Şanlı[1], Ç. Sıcakyüz[2]*, O.H. Yüregir[3]

[1] Department of Industrial Engineering, Çukurova University, TURKEY
[2] Department of Industrial Engineering, Ankara Bilim University, TURKEY*
[3] Department of Industrial Engineering, Çukurova University, TURKEY
[*]Corresponding Author: e-mail: cigdem.sicakyuz@ankarabilim.edu.tr, Tel +90-537-7910472
ORCID iD: http:/orcid.org/0000-0002-6706-3423 (Şanlı), https://orcid.org/0000-0002-1076-7980 (Sıcakyüz), https://orcid.org/0000-0002-9607-8149 (Yüregir)

Abstract

Data mining, which has different uses such as text mining and web mining, is especially used for clustering and classification purposes. In this study, this method was used for both classification and text mining. The aim of the study was the assessment of the performances of the data mining algorithms on the three datasets. A total of 6631 master's and doctoral dissertations written in the field of industrial engineering were downloaded from the Higher Education Council database. With the help of summary, subject titles and keywords of these dissertations, it was tried to be guessed which sub-field of industrial engineering it belongs to using WEKA program. As a result, it was observed that the data set containing the keywords obtained by weighting the expert opinion was more successful than the other two data sets. And the three most successful classification algorithms were found to be kNN, SMO, and J48, respectively.

Keywords: Classification Algorithms, Data Mining, Multiple Classes, Dataset.

DOI: http://dx.doi.org/10.4314/ijest.v12i3.8

Cite this article as:
Şanlı T., Sıcakyüz Ç., Yüregir O.H. 2020. Comparison of the accuracy of classification algorithms on three data-sets in data mining: Example of 20 classes. *International Journal of Engineering, Science and Technology*, Vol. 12, No. 3, pp. 81-89. doi: 10.4314/ijest.v12i3.8

Received: July 8, 2020; Accepted: August 3, 2020; Final acceptance in revised form: August 28, 2020

1. Introduction

It is a widely accepted fact that the importance of data mining is increasing day by day and this technique is used in different sectors and for various purposes such as making predictions in the financial field, diagnosing diseases in the field of health, determining the credit card thefts in online payments, and identifying the target audience in the field of marketing (Patil & Sherekar, 2013). Additionally, data mining has been useful in the field of education (Kabakchieva, 2013). Since the main purpose of data mining is to extract meaningful information from a data stack, in this way, the methods of estimating, defining, and establishing the association rule from data mining are utilized.

In this study, the classification method of data mining is discussed. As can be understood from its name, “classification” is to divide the available data in line with the purpose of the study and to separate them into new categories. Classification methods are made with different algorithms. The most widely used algorithms in the literature are as follows: Decision Tree, Artificial Neural Network (ANN), Support Vector Machine (SVM), Logistic Regression (LR), Discriminate Analysis (DA), Rule Based System and Bayesian Belief Networks and, Multilayer Perceptron Classifier (MLP), Sequential Minimal Optimization (SMO), J48. According to the literature, the performance of these algorithms varies and it is observed that different algorithms have different

results on different data sets (Labib & Rayed, 2020). Traditionally, the evaluations of algorithms in terms of space and time are in the secondary plan (Patil & Sherekar, 2013). Instead, the number of classes that the classifying algorithms classify correctly is more important. Because the correct classification rates of algorithms are taken into consideration in the literature. However, there is confusion as the evaluations are based on users. For example, decision trees are preferred instead of using neural network algorithms in classification. In some cases, neural networks give lower classification errors than decision trees. However, it is seen that neural networks require more time for training (Arora & Suman, 2012).

In the studies conducted, it is seen that the performance of the classifiers differs on the same data set in the classification process. The reasons that algorithms perform differently in different studies are the quality and quantity of the data set in which they are used. For example, according to (Dogan & Tanrikulu, 2013), the success rate of the classifier was affected by all features of the data set and applications such as PCA (principal component analysis). However, discretization did not have an impact on the success rate of classification.

Minaei-Bidgoli et al. (2003) observed the following in their studies using data mining method in the field of web-based learning: They observed that the performances of different algorithms changed according to the number of 2, 3 and 9 classes (Quadratic Bayesian classifier nearest neighbor (I-NN), K-Nearest Neighbor (K-NN), Parzen-Window, Multilayer Perceptron (MLP), and Decision Tree). It was revealed that as the number of classes increased, the performance rates of the algorithms decreased. This situation is called “multi-class problem” in the literature and recently, there is a growing attention on that problem for precise labelling of the groups especially as the number of classes increases (Singh & Singh, 2019). In case of pattern recognition such as face and image detection or finger pointing identification, multi-class problem has been faced (Rocha and Goldenstein, 2014). It has crucial importance on making decisions in many situations due to the inevitable results which can have some harmful effects on humans such as in the early detected stages of cancer. In the correct diagnosis of a disease, it is vital to make a decision based on the available data. Unfortunately, wrong analysis may lead to irreversible wrong decisions.

In diagnosing the presence of some diseases, certain distinct features (for example, the presence of a substance in the blood) may be sufficient to recognize that disease, while a few features may not be sufficient to determine the types of some diseases, and in this case, it is necessary to look at many features in order to decide which type of the disease in question. It may be necessary to evaluate. Therefore, more features may be required for the correct decision of different number of disease types, but in this case, decision may be more difficult. Thus, it is important to put the attributes in the right class of disease so that the diagnosis would be correct. The used classifier should predict the sorts of diseases in the right way regardless of the high number of classes. Some weak single classifier such as Linear Discriminant Analysis does not perform well on multi-classes classification without hybrid method (Rocha and Goldenstein, 2014). On the other hand, the complexity of the classification technique can rise because training samples can have many redundant and noisy data which has a negative effect on the quality of data (Singh & Singh, 2019). That is why it is important to choose the right classifier for diagnosing a disease as well as in other fields. This study, in this context, gives a perspective on the performance of the above mentioned classifiers that run under multi-class classification problems.

The purpose of this study is to observe how the performance of different algorithms changes according to different data sets on the same subject. The data of this research is taken from the study that determined the most appropriate category among the 20 categories previously determined by experts according to the subject, summary and keywords of theses in the field of industrial engineering.

In this study, kNN, J48, SMO and Naive Bayes, NBM, BAGGING and JRIP are selected as classification algorithms. The difference and specificity of the studies in the field of text mining vary according to the study. However, the type, size, and method of preparing the data set which is the most important factor in such studies can change the results of the study. The fact that the data sets are different, real and large reveals the importance of this study. For this reason, it guides the researchers working in the field of data mining in this direction. Also, knowing which algorithm performs best will make it easier for the researchers to choose. When algorithms are compared in the studies on data mining in general, while taking into consideration the data sets and algorithm types, very few studies have been encountered by considering the high number of classes with a great data. Having class number as 20 in this study also adds originality to this study.

2. Classification Algorithms

2.1 kNN (k-nearest neighbors classifier): The k-Nearest Neighbor algorithm is a supervised algorithm and is used in statistical prediction method and pattern recognition. The goal of this algorithm is to classify the objects according to the majority of the neighbors closest to it. In this model space, k is a positive integer indicating the number of neighbors and can never be larger than the data set (Arbain & Balakrishnan, 2019). When k = 1, unknown samples in the model space are assigned to the class of the training sample closest to it (Rajamohana et al., 2018). The accuracy of the kNN algorithm is influenced by the magnitude of k because the large value of k reduces the effect of the noise variable in the classification and makes the boundaries between the classes less visible (Kabakchieva, 2013).

2.2 Sequential minimal optimization (SMO): SMO is a new, fast, and easy algorithm proposed for training Support Vector machines (SVMs). And the purpose of this algorithm is to generate a quadratic optimization problem to solve. SMO requires a series of small quadratic programming problems that differ from large quadratic problems (Rajamohana et al., 2018).

2.3 Naive Bayes classifier: Bayesian classifiers are popular classification algorithms because of their ease, efficiency in computer, fast training and success in real world problems and high accuracy in many areas (Kabakchieva, 2013). The Naive Bayes algorithm is a statistical classification technique and its classes take into account the possibilities they belong to. This classifier is based on calculating frequencies with a series of possibilities on the given data set. Naive Bayes classification algorithm is based on total probability and Bayes theorem. Theoretically, several Naive Bayes algorithms have been developed.

2.4 Decision tree algorithm J48: Decision trees are powerful classification algorithms (Menaka & Kesavaraj, 2019) defining the relationships between qualities and the relative importance of quality (Kabakchieva, 2013). These algorithms are advantageous because of the easy understanding and interpretation of the displayed rules and complex data preparation is not required. Also, these algorithms perform well in numerical and categorical data (Kabakchieva, 2013).

2.5 Rule learners: Two classifiers are considered in this algorithm. OneR is a single-level decision tree, all expressed in a set of rules that test a certain quality. It is a simple and inexpensive method and generally produces good rules with high accuracy in defining the structure of the data. This classifier is based on comparison with others. And it shows the predictive power of certain qualities. The Jrip classifier uses the RIPPER (Repeated Incremental Pruning to Produce Error Reduction) algorithm (Kabakchieva, 2013).

2.6 Bagging: Bagging is a holistic method used to improve the accuracy of the algorithm. Bagging parallels the approach with multiple classifiers in estimation. The result of each class is passed through a selection process (Rajamohana et al., 2018).

2.7 Random forest: Defined as a decision trees forest consisting of random and different tree loaded algorithms, this algorithm creates a collection of methods that make up a whole. It is best evaluated from multiple decision trees and chosen by the majority. It is considered one of the most powerful algorithms. It shows high performance in both classification and regression, but over-fitting is the main problem of this algorithm (Arbain & Balakrishnan, 2019).

3. Review of Literature

There are different results in terms of the performance of classification algorithms in data mining. For example, (Patil & Sherekar, 2013) compared algorithms using a confusing matrix on a dataset to evaluate the accuracy performance of Naive Bayes and J48 algorithms. As a result, they found that J48 showed better results (Kabakchieva, 2013). On the other hand, they categorized the students in five classes in order to examine their performance before and during the university education period (Kabakchieva, 2013). They utilized the Common Decision Tree Algorithm C4.5 (J48), NaiveBayes, BayesNet, Nearest Neighbor Algorithm (kNN), and two rule learners (OneR and JRip) algorithms for this. According to the author, the best performing algorithm is J48, followed by the JRip and kNN classifier while the lowest one was Bayesian algorithms with the accuracy rates below 70%.

Arbain & Balakrishnan (2019) compared Logistic Regression, Random Forest, K-Nearest Neighbor (kNN), and Artificial Neural Network algorithms that they used to predict liver disease and they observed that the algorithm with the highest accuracy rate was the kNN algorithm. They also found that the Random Forest algorithm was not suitable for their own work, although its performance seemed appropriate. Rajamohana et al. (2018) used Naive Bayes, Random Forest, Bagging, and Multiboost classification algorithms for the early diagnosis of ASD disease. They determined that the algorithm reaching the highest accuracy rate was Multiboost with a 93.18% accuracy rate. These authors also benefited from the SMO and IBK (K-Nearest Neighbors Classifier) algorithms in choosing the most appropriate recommendation system to be developed against two types of tumors (benign or malignant) in breast cancer and to support doctors in their decision making. As a result, they decided that the algorithm with the best performance was SMO (Arora & Suman, 2012), on the other hand, compared J48 and MLP algorithms in five different data sets and sample sizes, and found that MLP was the best performer on all five data sets.

Kaya Keleş (2019), moreover, utilized the classification algorithms of data mining through the antenna information of cancer disease to determine whether the tumor existed. The study resulted that five top algorithms of data mining were Random Forest, Random Committee, Bagging, SimpleCART, and IBK, respectively. In a study that aimed to predict the diagnosis of Chronic Kidney Disease (CKD) according to their symptomps whether it was acute or chronic, they conducted the following algorithms to classify the CKD: ZeroR, Rule Induction, SVM, Naïve Bayes, Decision Tree, Decision Stump, k-NN, and Regression. They found that all of the classifiers had more than 90% accuracy rate apart from ZeroR and the best classifier was regression (Saringat et al., 2019). Arboleda (2019) used the 22 classification algorithms to sort four attributes of green coffee beans into three sort of its namely liberica, robusta, and excelsa. The result of the study showed that the Coarse Tree Algorithm (Coarse kNN) performed the best algorithm with the accuracy rate of 94.1 percent and 18 of 22 algorithms showed more than 90 percent accuracy. (Arboleda, 2019) also stated that it was worth to examine the data mining algorithms under a larger number of samples for verifying the relationship between the data size and accurate classification.

Mohammadi et al. (2020) ran five data mining algorithms (ANN, Bayesian Network, DA, LR, and SVM) to classify the companies in two groups (fraudulent and non-fraudulent companies) in their study about detecting financial statement fraud and

according to them, the best detecting algorithms was artificial neural network among the others. Labib and Rayed (2020) aimed to detect the most common type of cancer in childhood called Leukemia, in Egypt. For that, they used the three data mining algorithms (Decision Tree, Naïve Bayes, and Random Forest) in different classes to find the main risk factors (such as demographic, social, lifestyle, and environmental factors) of that disease. The class number was 18 and the result revealed that the most accurate algorithm was the decision tree.

Riri et al. (2020) used the classification algorithms on 1207 images of 98 different patients for recognition of orthodontic images. Firstly, they classified sixteen classes of orthodontic images such as extra-oral, mould, and intra-oral images by extracting features. For each image of three, they used one algorithm. Then, they merged the algorithms used to see the whole picture of all classes of orthodontic images. They used the Local Binary Pattern (LBP) to gain information and classified LBP with the classifiers Quadratic SVM, Cubic SVM, Radial Basis Function (RBF) SVM, Cosine k-NN, Euclidian kNN, and LDA. They finally implemented the principal component analysis (PCA) algorithm for optimization of the noisy parameters. Apart from Euclidian kNN, they found that the accuracy of the remained classifiers was high.

4. Method

4.1 Determination of sample and sample number: The sample of the study consists of 6631 industrial engineering thesis and dissertations, which were added to the Higher Education Council database between 1975 and 2018. Considering that the number of samples is infinite;

$$n_0 = \frac{(z^2 * p * q)}{e^2} \tag{1}$$

where n_0 = Number of samples, z = 1.96 (95% Confidence interval), P = Community ratio and p = 1-q

Formula for 10% sample error in 95% confidence interval;

$$\frac{1{,}96^2 * 0{,}5^2}{0{,}1^2} = 96$$

If the sample is in finite number,

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}} \tag{2}$$

where n = Number of samples where the sample space is finite, n_0= Number of samples and N = Total number of samples

$$94 = \frac{96}{1 + \frac{96 - 1}{6631}}$$

As a result, it is sufficient to know which topic and/or topic titles of 94 dissertations are under the correct functioning of the system. In this study, 6631 dissertations were downloaded and 400 of these dissertations were considered as training sets. For these 400 dissertations, datasets consisting of a sufficient amount of dissertations were prepared in three different ways by determining which topic and/or subject titles to be under by referring to expert opinions.

4.2 Determination of the word vector space: Text mining was used to determine the word vector space. In this study, three separate word vector spaces were prepared for three sets of training data. The objects in vector spaces were defined in the vector structure and the properties of these objects formed the axes of the vector space. Thanks to the positions of these vectors, the proximity of objects to each other could be calculated. Different methods were used in the preparation of the vector space model. For example, it can be seen that it is critical that a word must pass in a certain document, and it is considered important that the word fully complies with the subject of that document. According to the researches, it was stated that TF-IDF method was the most effective method in creating word space (Noh et al., 2015). This method measures the state of the word being related to the relevant document, that is, the determination and frequency of the word. TF shows the term frequency, that is, how many times the term has passed in the relevant document. However, TF cannot measure the words in the document that are not related to the high

frequency document. Therefore, the concept of IDF (inverse document frequency) was not introduced. And this dealt with the frequency of terms that were rarely found in all documents in the total (Usui et al., 2007). For this reason, this method was used directly in the preparation of the training sets of this study. However, the expert opinion of the V1 training set was also effective in weighting the words. For example, the frequency value of a word indicates the importance of that word. For example, if a keyword never exists in the summary, the importance of that word is considered to be zero. This takes a $2^0 = 1$ weighted value. However, if it has some degree of significance, the significance level of this word is taken as $2^1 = 2$. If the keyword is two words, then the significance of this keyword is $2^2 = 4$. If the expert opinion argues that this keyword is suitable for this thesis, the significance of this word is determined as $2^3 = 8$. The importance of expert opinion in weighting the keywords made a difference in this way. Other details of the study's datasets are given below.

4.2.1 First training kit V1: Expert opinion was used in this training set. On the text mining studies in the literature, it can be seen that, in general, keywords are decided according to whether the text belongs to a predetermined class or classes. And this is the final result of text mining. However, in this study, these classes determined with text-mining will be reclassified later with classification algorithms. Therefore, it is seen as a pre-processing step in determining the correct categories before classification. The reason for this can be determined in which category of industrial engineering will be evaluated with the help of experts, topics, summary, and keywords of thesis studies. Thus, it is thought that better results can be obtained in cases where the machine will be insufficient. For example, a word can be included in more than one workspace. Therefore, the experience and intuition of the expert industrial engineer are vital in this case. Only in this way, it is possible to evaluate the working areas in an integrity.

For this purpose, a web application was prepared and 10% of the dissertations downloaded were presented to expert opinions. The data obtained here were also used in the system as a training data set. And it was also listed in the subject headings in the prepared web application. The words and/or word groups entered into the system were included in all three data sets as expert suggestions of these dissertations. A platform called “Thesis Portal” was created in order to reflect the expert opinions about the dissertations that were registered to the program more easily. In this way, it was aimed to contribute to the database by receiving the opinions of the experts on the subject and keywords based on the thesis content.

In the preparation of training sets, vector space was positioned as the training set, control set, and prediction set. As vectors of words from the first training set (V1) [3156,400], [3156,1600], [3156,6631] 3 files with the extension ".arff" were created.

4.2.2 Second training set V2: This training set consists of keywords included in thesis, determined by students and academics. As vectors of words from this training set (V2) [3582,400], [3582,1600], [3582,6631] 3 files with the extension ".arff" were created.

4.2.3 Third training set V3: This training set includes all the topics of the thesis title, thesis summary, thesis keywords. The data in this training set were determined according to the frequency and weight scores of the words. As vectors of words from the third training set (V3) [5272,400], [5272,1600], [5272,6631] 3 files with the extension ".arff" were created.

4.2 Methodology: A total of 6631 master's and doctoral dissertations, which were written under the umbrella of industrial engineering between 1975-2018, were downloaded from the relevant data source. And they underwent data cleaning before the classification. These operations can be listed as eliminating the stop words, cleaning the spaces in the word and number and punctuation, reducing the words to the roots in line with the Zemberek library.

Moreover, the dissertations included in the study were classified with the trained “Naive Bayes” algorithm. Then, these data were analyzed on all 3 data sets of dissertations using "BAGGING", "J48", "JRIP", "kNN", "NB", "NBM", "SMO" algorithms that come as a package in the "Weka" program. Based on this, 21 different classification results were transferred to the database. At the stage of testing the validity of the algorithm on the control set, the dissertations classified were compared with the control set. As a result of the comparison process, recall values were calculated and stored in the system. In addition to the results transferred to the database, the operating times and accuracy rates of the algorithms were also recorded. The results obtained in the "Weka" program achieved a coefficient in direct proportion to the accuracy rates of each algorithm. And by taking the weighted average of these coefficients, the classes of dissertations were determined.

5. Results and discussion

As a result of the classification, the subject titles of the dissertations in the field of industrial engineering were estimated and compared with each other according to the data sets created.

5.1 Comparing data sets: Analyzes were made on 3 different data sets. The graphic presented in Figure 1 was prepared according to the average of the results of 8 algorithms that operate on data sets.

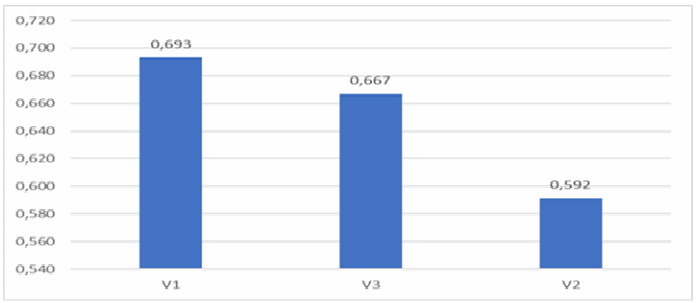

Figure 1. Average accuracy results by data sets

Figure 1 shows the average accuracy results of the data sets according to the algorithms used. As can be seen from the graphic, the V1 dataset gave the most accurate result in all algorithms run on average. The V3 dataset also showed almost the same performance as the V1 training set. In addition, when the results are examined, it is seen that the data set that has the least impact on the algorithms is V2. One reason for the low ranking of the V2 data set may be the keywords that the students who prepared the thesis study stated in their thesis studies. Because the technical terms written in these keyword fields may show specific fields, possibly preventing the algorithms being run from simulating.

The datasets were also compared in terms of speed and the average time graph in analyzing the datasets given in Figure 2 below, in seconds (s).

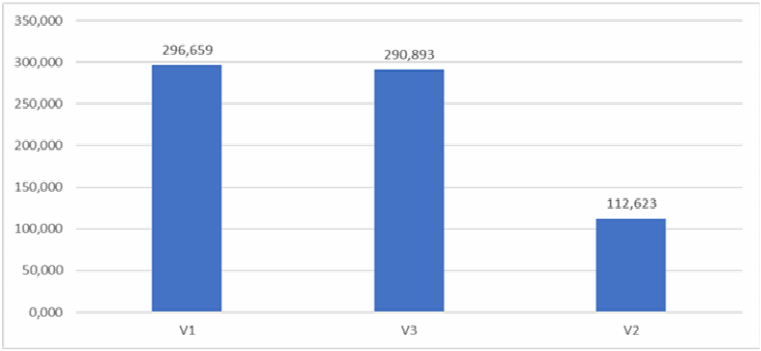

Figure 2. Average time graph by data sets

As seen in the graphic in Figure 2, when the data sets on which the algorithms are run were examined on a time basis, the data set that requires the most time was the V1 data set with 296.65 seconds, while the data set that required the least time was the V2 data set with 112.62 seconds. Reasons for this include that V1 and V3 datasets were higher in number than V2 datasets in terms of vector space. Another possible reason for the difference that can be seen on the results was due to the fact that the algorithms were run one after the other on the datasets, creating a fluctuating effect on the times in terms of both processor and memory density.

5.2 Comparison of algorithms: In this study, the accuracy rates of the algorithms run on each data set are shown in Table 1 below. As a result of the study, 7 algorithms were run on each data set and 21 different results were obtained. However, COUNT and SUM algorithms were calculated by taking the average of all 21 different results. The item shown with "V123" in Table 1 means that all three data sets were analyzed together. By taking the arithmetic average of these 21 results, COUNT and the geometric average of the SUM algorithm were created. In this way, for example, the effect of the accuracy rates of each result was reflected in the SUM algorithm. Apart from these algorithms, a Bayes-based algorithm named MYNaiveBayes (MYNB) was also developed to be compared with other Bayes algorithms.

Table 1. Accuracy rates of algorithms on datasets

	BAGGING	J48	JRIP	kNN	MYNB	NB	NBM	SMO	SUM	COUNT	AVERAGE
V123									0,82	0,82	0,82
V1	0,66	0,75	0,56	0,84	0,51	0,71	0,68	0,84			0,69
V3	0,66	0,75	0,53	0,84	0,34	0,71	0,68	0,84			0,67
V2	0,56	0,63	0,32	0,83	0,59	0,49	0,59	0,72			0,59
AVERAGE	0,63	0,71	0,47	0,84	0,48	0,64	0,65	0,80	0,82	0,82	0,65

As can be seen from Table 1, in this study, it was observed that the algorithms run on V1 and V3 datasets gave close results. In addition, the results of the algorithms run on the V2 dataset showed lower results than the other two datasets. Also, when the averages of success are taken into account on 3 different data sets, it is seen that the kNN algorithm took the first place with a success of 0.83. At the same time, this algorithm had the highest success performance on all three data sets.

On average, algorithms with second and third performance on three data sets were observed as SUM and COUNT. This means that the algorithms produced substantially the same correct results on any thesis. It is seen that the most successful algorithms except for the kNN algorithm, where the performance of performance shown on each data set is highest, are SMO and J48, respectively. In the literature, it is seen that J48 algorithm is more successful than many algorithms. However, in this study, the SMO algorithm performed more than J48. One reason for this may be the size of the data available. Because for (Rajamohana et al., 2018) since the amount of space required for the training set is linear, SMO can deal with large data sets. In various test problems, SMO is somewhere between linear and quadratic in the training set size because it avoids matrix calculations and SMO is the fastest algorithm for linear SVMs and sparse data sets.

Another result seen in the table is that the “Naive Bayes” algorithm and/or “Naive Bayes” based algorithms performed between 48% and 66%. The algorithm named MYNB, on the other hand, showed a calculated "Naive Bayes" algorithm in this study. In addition, it is seen that the “JRIP” algorithm, which was run on the V2 algorithm with the accuracy rate of approximately 32%, showed the lowest performance. While the accuracy rate of the NB algorithm, which is used frequently in literature research, was about 64%, the average accuracy rate of the MNB algorithm was calculated as 65%. The reasons for this can be based on the studies of Altintas (2014) and Reiten (2017). For them, algorithms such as “Naive Bayes” give approximately the same results on datasets that can be called insufficient, because, as a result of the operation of the related algorithm, there is an increase and decrease in the rates by returning the similarity rate as the output and this affects the result of the algorithm in certain rates.

The results of this study differ from the study of (Kabakchieva, 2013) because in her study, J48 algorithm was more successful than kNN. However, in this study and in the study of (Arbain & Balakrishnan, 2019) kNN performed as the best. One reason for this difference can be related to the size of the data set used because the J48 algorithm performed well on large data sets, especially when the number of attributes was high, the tree would grow bigger and would require a lot of time for calculation (Ozer, 2008). In addition, this study is similar to that of (Rajamohana et al., 2018). According to them, SMO and kNN showed the best performance. From this point of view, SMO can be a good alternative when it comes to big data sets. The kNN algorithm performs better than the INN algorithm when the number of classes is two on the same data set. When the number of classes was 9 (Kaya Keleş, 2019), it performed lower than this algorithm. Similarly, when the number of classes was two, the Bayes algorithm and INN algorithm showed almost par with performance. However, when the number of classes was 9 (Kaya Keleş, 2019), Bayes lagged far behind in terms of performance (Minaei-Bidgoli et al., 2003).

This results diverges from the study of (Horak et al., 2017). Their study about the detection of license plate in gallery with 535 images of different vehicles resulted that the Naïve Bayes Multivariate algorithm had high accuracy of the value of 99.8%. kNN is a nonparametric classification method that is theoretically not based on a mathematical density function model (Güney & Atasoy, 2012) and even though it could show bias variance when the sample size is limited, (Zhang et al., 2006) listed some advantages of the kNN. One of them is that it does not require certain structure of an attribute space and is capable of coping with highly multiclass nature of visual object recognition easily. Additionally, as the sample size approaches to infinity, the error rate of kNN treats as Bayes optimal classifier. The sample size plays role in the correctness of decision for the defining classes. In the case of multi- class decision, the classification method could be more important than the binary ones. As for kNN, the performance of multi label-kNN differ from various values of k (number of neighbors). To handle multi-class classification problems, some authors suggested to extend the algorithms (Zhang & Zhou, 2005) or hybrid models (Zhang et al., 2006), or feature manipulation technique (Jia & Zhang, 2020) to give better solution for prediction. Additionally, there are some studies that show that kNN based on the decision tree structure increased the success rate of kNN (Güney & Atasoy, 2012). In this study, kNN algorithm individual performed well on different but sufficient size datasets for categorizing 20 classes with the success rate of 84 %, while (Güney & Atasoy, 2012) achieved the success rate 96% of kNN with decision structure on the insufficient data size. As a result, the high number of classes could increase the sensitivity of the algorithm and more accurate results can be encountered.

6. Conclusions

In this study, seven different algorithms were run on three different data sets. It was revealed that the most successful data set was V1, that is, the data set in which expert opinions were taken. It was observed that the V3 data set was also very successful compared to the V 'set. From this point of view, getting expert opinion as a support for machine learning may have more accurate results. Another result of this study is that kNN was the most successful algorithm among the related classification algorithms. And following this, it is seen that the SMO and J48 algorithms ranked. The absence of K-Fold Cross-Validation was a limitation of this study. Because of the large size of the data set and the number of classes, this verification method was not found suitable because it slowed down the system and occupied a lot of space. In addition, the high number of classes can cause accuracy rates to reduce. When the number of classes is less, it can be thought that these algorithms studied will show more accuracy. In the study, the word count of all thesis abstracts was not taken into consideration. In later studies, this study should be done by taking into

account the length of the abstracts and the accuracy rates should be compared. The next study should run different algorithms on a similar subject with a higher number of classes and more data, and interpret these algorithms by comparing them.

References

Altıntaş M. 2014. Kullanıcı destek sistemlerinde yardım biletlerinin otomatik sınıflandırılması [Doctoral dissertation]. Turkey: Çukurova University.

Arbain, A. N., & Balakrishnan, B. Y. P. 2019. A comparison of data mining algorithms for liver disease prediction on imbalanced data. *International Journal of Data Science and Advanced Analytics*, Vol. 1, No. 1, pp. 1-11.

Arboleda, E. R. 2019. Comparing performances of data mining algorithms for classification of green coffee beans. *International Journal of Engineering and Advanced Technology*, Vol. 8, No. 5, pp. 1563–1567.

Arora, R., & Suman, S. 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. *International Journal of Computer Applications*, Vol. 54, No. 13, pp. 21–25. https://doi.org/10.5120/8626-2492

Dogan, N., & Tanrikulu, Z. 2013. A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness. *Information Technology and Management*, Vol. 14, No. 2, pp. 105–124. https://doi.org/10.1007/s10799-012-0135-8

Güney, S., & Atasoy, A. 2012. Multiclass classification of n-butanol concentrations with k-nearest neighbor algorithm and support vector machine in an electronic nose. *Sensors and Actuators, B: Chemical*, Vol. 166–167, pp. 721–725. https://doi.org/10.1016/j.snb.2012.03.047

Horak, K., Klecka, J., Bostik, O., & Davidek, D. 2017. Classification of SURF image features by selected machine learning algorithms. *2017 40th International Conference on Telecommunications and Signal Processing, TSP 2017*, *2017-Janua*(July), pp. 636–641. https://doi.org/10.1109/TSP.2017.8076064

Jia, B. Bin, & Zhang, M. L. 2020. Multi-dimensional classification via kNN feature augmentation. *Pattern Recognition*, Vol. 106. https://doi.org/10.1016/j.patcog.2020.107423

Kabakchieva, D. 2013. Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, Vol. 13, No. 1, pp. 61–72. https://doi.org/10.2478/cait-2013-0006

Kaya Keleş, M. 2019. Breast cancer prediction and detection using data mining classification algorithms: A comparative study. *Tehnicki Vjesnik*, Vol. 26, No. 1, pp. 149–155. https://doi.org/10.17559/TV-20180417102943

Labib, S. E., & Rayed, C. A. 2020. Prediction model for risk factors of childhood leukemia based on data mining classification algorithms. *Egyptian Computer Science Journal*, Vol. 44, No. 2, pp. 51–63.

Menaka, S., & Kesavaraj, G. 2019. Predicting student performance using data mining techniques: A survey of the last 5 years. *International Journal of Advanced Scientific Research and Management*, Vol. 4, No. 1, pp. 98-102.

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003). Predicting student performance: An application of data mining methods with an educational web-based system. *Proceedings - Frontiers in Education Conference, FIE*, *1*, T2A13-T2A18. https://doi.org/10.1109/FIE.2003.1263284

Mohammadi, M., Yazdani, S., Khanmohammadi, M. H., & Maham, K. 2020. Financial reporting fraud detection: An analysis of data mining algorithms. *International Journal of Finance & Managerial Accounting*, Vol. 4, No. 16, pp. 1–12.

Noh, H., Jo, Y., & Lee, S. 2015. Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications*, Vol. 42, No. 9, pp. 4348–4360. https://doi.org/10.1016/j.eswa.2015.01.050

Ozer, P. 2008. *Data Mining Algorithms for Classification*. *January*, 26. https://doi.org/doi=10.1.1.331.4778

Patil, T. R., & Sherekar, S. S. 2013. Performance analysis of ANN and naive bayes classification algorithm for data classification. *International Journal of Computer Science and Applications*, Vol. 6, No. 2. https://doi.org/10.18201/ijisae.2019252786

Rajamohana S. P., Dharani A., Anushree P., Santhiya B., & Umamaheswari K. 2018. *Machine Learning Techniques for Healthcare Applications*. pp. 236–251. https://doi.org/10.4018/978-1-5225-7522-1.ch012

Reiten T.E.G. 2017. Classification with multiple classes using Naïve Bayes and text generation with a small data set using a recurrent neural network [Master's thesis]: Universitetet i Agder.

Riri, H., Ed-Dhahraouy, M., Elmoutaouakkil, A., Beni-Hssane, A., & Bourzgui, F. 2020. Extracted features based multi-class classification of orthodontic images. *International Journal of Electrical and Computer Engineering*, Vol. 10, No. 4, pp. 3558–3567. https://doi.org/10.11591/ijece.v10i4.pp3558-3567

Rocha A.; Goldenstein S.K. 2014. Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches. *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 25, No. 2, pp. 289–302. DOI: 10.1109/TNNLS.2013.2274735

Saringat, Z., Mustapha, A., Saedudin, R. D. R., & Samsudin, N. A. 2019. Comparative analysis of classification algorithms for chronic kidney disease diagnosis. *Bulletin of Electrical Engineering and Informatics*, Vol. 8, No. 4, pp. 1496–1501. https://doi.org/10.11591/eei.v8i4.1621

Singh, N., & Singh, P. 2019. A novel bagged naïve bayes-decision tree approach for multi-class classification problems. *Journal of Intelligent and Fuzzy Systems*, Vol. 36, No. 3, pp. 2261–2271. https://doi.org/10.3233/JIFS-169937

Usui, S., Palmes, P., Nagata, K., Taniguchi, T., & Ueda, N. 2007. Keyword extraction, ranking, and organization for the neuroinformatics platform. *BioSystems*, Vol. 88, No. 3, pp. 334–342. https://doi.org/10.1016/j.biosystems.2006.08.015

Zhang M.-L., & Zhou Z.-H. 2005. A k-nearest neighbor based algorithm for multi-label classification. *2005 IEEE International Conference on Granular Computing*, 25-27 July 2005, Vol. 2, pp. 718-721 https://doi.org/10.1109/grc.2005.1547385

Zhang, H., Berg, A. C., Maire, M., & Malik, J. 2006. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 2126–2136. https://doi.org/10.1109/CVPR.2006.301

Biographical notes

T. Şanlı received the Bachelor degree in Computer Engineering from Çukurova University, in 2013. He earned his Msc. degree in Industrial Engineering department of Çukurova University. Currently, he works as a computer engineer.

Ç. Sıcakyüz graduated from Industrial Engineering department of Osmangazi University, Eskişehir, in 2001. In 2009, she got her MSc degree in Production Engineering from Bremen University in Germany. She completed her Ph.D. at Industrial Engineering Department of Çukurova University in 2019. She works currently in the department of Industrial Engineering from the University of Ankara Bilim. Ankara, Turkey.

O.H. Yüregir is an Associate Professor at the Department of Industrial Engineering of Çukurova University, Adana, Turkey. She is a graduate of Boğaziçi University (AD), Anadolu University (BA), the University of Texas at Austin (MBA) and Çukurova University (Ph.D.). Her research interests include software engineering, information systems, process management, and innovation management.