International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

Comparison of Classification Techniques on Energy Efficiency Dataset

Ahmet TOPRAK[1], Nigmet KOKLU[2], Aysegul TOPRAK[3], Recai OZCAN[1]

Accepted : 13/06/2017 Published: 30/06/2017

Abstract: The definition of the data mining can be told as to extract information or knowledge from large volumes of data. Statistical and machine learning techniques are used for the determination of the models to be used for data mining predictions. Today, data mining is used in many different areas such as science and engineering, health, commerce, shopping, banking and finance, education and internet. This study make use of WEKA (Waikato Environment for Knowledge Analysis) to compare the different classification techniques on energy efficiency datasets. In this study 10 different Data Mining methods namely Bagging, Decorate, Rotation Forest, J48, NNge, K-Star, Naïve Bayes, Dagging, Bayes Net and JRip classification methods were applied on energy efficiency dataset that were taken from UCI Machine Learning Repository. When comparing the performances of algorithms it's been found that Rotation Forest has highest accuracy whereas Dagging had the worst accuracy.

Keywords: *Data mining, classifications, energy efficient.*

1. Introduction

Developments in Information Technology and database software immense amount of data are collected. This large amount of data has appeared as one of the culprits of meaningful knowledge extraction. Collected large amount of data although contains hidden patterns, as the amount of the data increases, cannot be converted into useful information by traditional methods. Consequently, to analyze the immense amount of data, fairly new method known as data mining methods are widespread in practice [1].

Data mining is used as an information source to find unities, make classification, clustering and estimations by using information discovery systems which are the combination of data warehouses, artificial intelligence techniques and statistical methods [2][3].

Classification is a method frequently used in data mining and used to uncover hidden patterns in database. Classification is used to insert the data object into predefined several classes. The well-defined characteristics play a key role in performance of the classifier. Classification is based on a learning algorithm. Training cannot be done by using all data. This is performed on a sample of data belonging to the data collection. The purpose of learning is the creation of a classification model. In other words classification is a class determination process for an unknown record [4][5][6].

Energy consumption of buildings has received increasing great interest in today's economies. As buildings represent substantial consumers of energy worldwide, with this trend increasing over the past few decades due to rising living standards, this issue has drawn considerable attention. The largest part of the energy consumption is due to the use of so-called heating, ventilation and air-conditioning systems in the residential buildings. High energy consumption of buildings and the increase in building energy demand require the design of energy efficient buildings and an improvement of their energy performance. One way to reduce the increased energy demand is to have more energy-efficient building designs. Another significant issue is the effect of this continuous increase of energy consumption on the environment. Buildings use about 40% of global energy, 25% of global water and 40% of global resources according to United Nations Environment Program (UNEP). There is a main danger that, as a consequence of global warming and climate change, energy demand and CO2 emissions will increase even further in many countries. In particular, the buildings design has a major impact on its energy footprint. In order to reduce the impact of building energy consumption on the environment, the European Union has adopted a directive requiring European countries to conform to proper minimum requirements regarding energy efficiency [7] [8].

Designing energy efficient buildings, it is important for architects, engineers and designers to identify which parameters will significantly influence future energy demand. After the identification of these parameters, architects and building designers usually need simple and reliable methods for rapidly estimating building energy performance, so that they can optimize their design plans. In recent years, several methods have been proposed for modeling building energy demand. For the estimation of the flow of energy and the performance of energy systems in buildings, analytic computer codes are often used [7][9].

2. Literature Survey

Tsanas and Xifara (2012) developed a statistical machine learning framework to study the effect of eight input variables (relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, glazing area distribution) on two output variables, namely heating load (HL) and cooling load (CL), of residential buildings. Extensive simulations on 768 diverse residential buildings show that they can predict HL and CL [10].

Castelli et all. (2015) proposed a genetic programming-based

[1]*Bozkir Vocational School, Department of Electric and Energy, Selcuk University, Konya, TURKEY*
[2]*Vocational School of Technical Sciences, Department of Construction, Selcuk University, Konya, TURKEY*
[3]*Kadinhani Faik Icil Vocational School, Department of Electronic and Automation, Selcuk University, Konya, TURKEY*
**Corresponding Author: Email: atoprak@selcuk.edu.tr*

framework for estimating the energy performance (the heating load and the cooling load) of residential buildings. The proposed framework blends a recently developed version of genetic programming with a local search method and linear scaling. The resulting system enables to build a model that produces an accurate estimation of both considered parameters. Extensive simulations on 768 diverse residential buildings confirm the suitability of the proposed method in predicting heating load and cooling load [7].

3. Material And Method

3.1. Dataset

The energy efficiency dataset used in this study was taken from UCI Machine Learning Repository. We perform energy analysis using 12 different building shapes simulated in Ecotect. The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, amongst other parameters. We simulate various settings as functions of the afore-mentioned characteristics to obtain 768 building shapes. The dataset comprises 768 samples and 8 features, aiming to predict two real valued responses. It can also be used as a multi-class classification problem if the response is rounded to the nearest integer [11].

In this study, we investigate the effect of eight input variables: Relative compactness (X1), surface area (X2), wall area (X3), roof area (X4), overall height (X5), orientation (X6), glazing area (X7), and glazing area distribution (X8), to determine the output variables Heating load (Y1) and Cooling load (Y2) of residential buildings. The dataset contains eight attributes (or features, denoted by X1...X8) and two responses (or outcomes, denoted by Y1 and Y2). The aim is to use the eight features to predict each of the two responses [10][11].

3.2. Software-WEKA

Weka (Waikato Environment for Knowledge Analysis) written in Java, developed at the University of Waikato, New Zealand [11]. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All techniques of Weka's software are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported) [12][13].

3.3. Methods

Having done in this study 10 different classifying techniques were used to energy efficiency. Short information about each of the classifying techniques namely Bagging, Decorate, Rotation Forest, J48, NNge, K-Star, Naïve Bayes, Dagging, Bayes Net and JRip will be mentioned in the following paragraphs.

Bagging (Bootstrap Aggregating) algorithm uses bootstrapping (equiprobable selection with replacement) on the training set to create many varied but overlapping new sets. The base algorithm is used to create a different base model instance for each bootstrap sample, and the ensemble output is the average of all base model outputs for a given input [14][15][16][17][18].

Decorate (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) iteratively generates an ensemble by learning a new classifier at each iteration. In the first iteration the base classifier is built from the given training data set and each successive classifier is built from an artificially generated training data set which is the result of the union of the original training data and artificial training examples, known as diversity data. The classifier built from the new training data set is added to the ensemble only if it reduces the ensemble training error, otherwise it is rejected and the algorithm continues iterating. Artificial training examples are generated from the data distribution and they are obtained by probabilistically estimating the value of each attribute. The labels for the new examples are selected with a probability that is inversely proportional to the prediction of the current ensemble. Decorate tries to maximize the diversity of the base classifiers by adding new artificial examples and re-weighting the training data [14][19].

Rotation Forest is a classifier that transforms the dataset to generate ensemble of classifiers. In this classifier, each base classifier is trained which extracts attributes in a different sets. The main goal is to embed feature extraction and reform approximately an attribute set for each classifier in the ensemble [20][21].

J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility [22].

NNge learns incrementally by first classifying and then generalizing each new example. It uses a modified Euclidean distance function that handles hyperrectangles, symbolic features, and exemplar and feature weights. Numeric feature values are normalized by dividing each value by the range of values observed. The class predicted is that of the single nearest neighbor. NNge uses dynamic feedback to adjust exemplar and feature weights after each new example is classified. When classifying an example, one or more hyperrectangles may be found that the new example is a member of, but which are of the wrong class. NNge prunes these so that the new example is no longer a member. Once classified, the new example is generalized by merging it with the nearest exemplar of the same class, which may be either a single example or a hyperrectangle. In the former case, NNge creates a new hyperrectangle, where as in the latter it grows the nearest neighbor to encompass the new example. Over generalization, caused by nesting or overlapping hyperrectangles, is not permitted. Before NNge generalizes a new example, it checks to see if there are any examples in the affected area of feature space that conflict with the proposed new hyperrectangle. If so, the generalization is aborted, and the example is stored verbatim [23].

K-Star algorithm is an instance-based classifier that uses entropic distance measurement with different data sets. It produces a predictive pattern by using some similar function. The class of a test instance is based on the training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function. Instance-based learners classify an instance by comparing it to a database of pre-classified examples. The fundamental assumption is that similar instances will have similar classifications. The question lies in how to define "similar instance" and "similar classification". The corresponding components of an instance-based learner are the distance function which determines how similar two instances are, and the classification function which specifies how instance similarities yield a final classification for the new instance. The K-star algorithm uses entropic measure, based on probability of transforming an instance into another by randomly choosing between all possible transformations [20][24].

Naïve Bayes algorithm is an intuitive method that uses the conditional probabilities of each attribute belonging to each class to make a prediction. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Parameter estimation for naive Bayes models uses the method of maximum likelihood. In spite over-simplified assumptions, it often performs better in many complex real world situations. One of the major advantages of Naive Bayes theorem is that it requires a small amount of training data to estimate the parameters [25][26].

Dagging is meta classifier creates a number of disjoint, stratified folds out of the data and feeds each chunk of data to a copy of the supplied base classifier. Predictions are made via majority vote [18][27].

Bayes Net is probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG) [11][18].

JRip (Java Repeated Incremental Pruning) is a prepositional rule learner, i.e. Repeated Incremental Pruning to Produce Error Reduction (RIPPER). Initial rule set for each class is generated using IREP [18][28].

4. Experimental Study

Bagging classifier technique was used to energy efficiency dataset and the results shown in Table 1 is obtained. Thus the correct classification ratio is % 65,8854 (Y1) and % 55,3385 (Y2).

Table 1. Accuracy Ratio of Bagging Application

Parameters	Value (Y1)	Value (Y2)
Correctly Classified Instances	506	425
Incorrectly Classified Instances	262	343
Kappa statistic	0,6405	0,5283
Mean absolute error	0,0225	0,0273
Root mean squared error	0,1072	0,1196
Relative absolute error	43,6243 %	54,5061 %
Root relative squared error	66,8524 %	75,6469 %
Total Number of Instances	768	768
Accuracy	65,8854 %	55,3385 %

Decorate classifier technique was used to energy efficiency dataset and the results shown in Table 2 is obtained. Thus the correct classification ratio is % 70,9635 (Y1) and % 57,0313 (Y2).

Table 2. Accuracy Ratio of Decorate Application

Parameters	Value (Y1)	Value (Y2)
Correctly Classified Instances	545	438
Incorrectly Classified Instances	223	330
Kappa statistic	0,6942	0,5470
Mean absolute error	0,0172	0,0252
Root mean squared error	0,1030	0,1268
Relative absolute error	33,4999 %	50,2608 %
Root relative squared error	64,2274 %	80,1791 %
Total Number of Instances	768	768
Accuracy	70,9635 %	57,0313 %

Rotation Forest classifier technique was used to energy efficiency dataset and the results shown in Table 3 is obtained. Thus the correct classification ratio is % 70,9635 (Y1) and % 58,9844 (Y2).

Table 3. Accuracy Ratio of Rotation Forest Application

Parameters	Value (Y1)	Value (Y2)
Correctly Classified Instances	545	453
Incorrectly Classified Instances	223	315
Kappa statistic	0,6941	0,5673
Mean absolute error	0,0192	0,0252
Root mean squared error	0,1023	0,1214
Relative absolute error	37,2073 %	50,3576 %
Root relative squared error	63,7658 %	76,7541 %
Total Number of Instances	768	768
Accuracy	70,9635 %	58,9844 %

J48 classifier technique was used to energy efficiency dataset and the results shown in Table 4 is obtained. Thus the correct classification ratio is % 70,5729 (Y1) and % 57,2917 (Y2).

Table 4. Accuracy Ratio of J48 Application

Parameters	Value (Y1)	Value (Y2)
Correctly Classified Instances	542	440
Incorrectly Classified Instances	226	328
Kappa statistic	0,6900	0,5494
Mean absolute error	0,0170	0,0252
Root mean squared error	0,1040	0,1296
Relative absolute error	33,0018 %	50,3089 %
Root relative squared error	64,8433 %	81,9644 %
Total Number of Instances	768	768
Accuracy	70,5729 %	57,2917 %

NNge classifier technique was used to energy efficiency dataset and the results shown in Table 5 is obtained. Thus the correct classification ratio is % 64,4531 (Y1) and % 56,1198 (Y2).

Table 5. Accuracy Ratio of NNge Application

Parameters	Value (Y1)	Value (Y2)
Correctly Classified Instances	495	431
Incorrectly Classified Instances	273	337
Kappa statistic	0,6263	0,5379
Mean absolute error	0,0192	0,0231
Root mean squared error	0,1386	0,1520
Relative absolute error	37,3160 %	46,1331 %
Root relative squared error	86,4314 %	96,1079 %
Total Number of Instances	768	768
Accuracy	64,4531 %	56,1198 %

KStar classifier technique was used to energy efficiency dataset and the results shown in Table 6 is obtained. Thus the correct classification ratio is % 63,4115 (Y1) and % 56,1198 (Y2).

Table 6. Accuracy Ratio of KStar Application

Parameters	Value (Y1)	Value (Y2)
Correctly Classified Instances	487	431
Incorrectly Classified Instances	281	337
Kappa statistic	0,6148	0,5375
Mean absolute error	0,0321	0,0336
Root mean squared error	0,1206	0,1267
Relative absolute error	62,251 %	67,1594 %
Root relative squared error	75,2038 %	80,0962 %
Total Number of Instances	768	768
Accuracy	63,4115 %	56,1198 %

Naive Bayes classifier technique was used to energy efficiency dataset and the results shown in Table 7 is obtained. Thus the correct classification ratio is % 36,7188 (Y1) and % 44,6615 (Y2).

Table 7. Accuracy Ratio of Naïve Bayes Application

Parameters	Value (Y1)	Value (Y2)
Correctly Classified Instances	282	343
Incorrectly Classified Instances	486	425
Kappa statistic	0,3357	0,4203
Mean absolute error	0,0368	0,0335
Root mean squared error	0,1491	0,1412
Relative absolute error	71,4952 %	66,9700 %
Root relative squared error	92,9719 %	89,2815 %
Total Number of Instances	768	768
Accuracy	36,7188 %	44,6615 %

Dagging classifier technique was used to energy efficiency dataset and the results shown in Table 8 is obtained. Thus the correct classification ratio is % 26,0417 (Y1) and % 29,9479 (Y2).

Table 8. Accuracy Ratio of Dagging Application

Parameters	Value (Y1)	Value (Y2)
Correctly Classified Instances	200	230
Incorrectly Classified Instances	568	538
Kappa statistic	0,2041	0,2412
Mean absolute error	0,0515	0,0502
Root mean squared error	0,1594	0,1575
Relative absolute error	99,9230 %	99,2806 %
Root relative squared error	99,3870 %	99,5951 %
Total Number of Instances	768	768
Accuracy	26,0417 %	29,9479 %

Bayes Net classifier technique was used to energy efficiency dataset and the results shown in Table 9 is obtained. Thus the correct classification ratio is % 53,7760 (Y1) and % 51,5625 (Y2).

Table 9. Accuracy Ratio of Bayes Net Application

Parameters	Value (Y1)	Value (Y2)
Correctly Classified Instances	413	396
Incorrectly Classified Instances	355	372
Kappa statistic	0,5130	0,4889
Mean absolute error	0,0329	0,0315
Root mean squared error	0,1288	0,1293
Relative absolute error	63,8258 %	62,8570 %
Root relative squared error	80,2829 %	81,7511 %
Total Number of Instances	768	768
Accuracy	53,7760 %	51,5625 %

JRip classifier technique was used to energy efficiency dataset and the results shown in Table 10 is obtained. Thus the correct classification ratio is % 58,2031 (Y1) and % 50,3906 (Y2).

Table 10. Accuracy Ratio of JRip Application

Parameters	Value (Y1)	Value (Y2)
Correctly Classified Instances	447	387
Incorrectly Classified Instances	321	381
Kappa statistic	0,5535	0,4633
Mean absolute error	0,0265	0,0312
Root mean squared error	0,1208	0,1280
Relative absolute error	51,3868 %	62,3570 %
Root relative squared error	75,3489 %	80,9383 %
Total Number of Instances	768	768
Accuracy	58,2031 %	50,3906 %

5. Results and Discussion

Following classifier techniques of WEKA have been applied to energy efficiency datasets: Rotation Forest, Decorate, J48, Bagging, NNge, KStar, JRip, Bayes Net, Naïve Bayes and Dagging, The results obtained from related classification techniques were presented in Table 11 according to each dataset, When comparing the performances of algorithms it's been found that Rotation Forest has highest accuracy whereas Dagging had the worst accuracy.

Table 11. Ratio Of Each Classification Technique On Each Energy Efficiency Dataset

No	Algorithm	Accuracy (%) Y1	Accuracy (%) Y2
1	Rotation Forest	70,9635	58,9844
2	Decorate	70,9635	57,0313
3	J48	70,5729	57,2917
4	Bagging	65,8854	55,3385
5	NNge	64,4531	56,1198
6	KStar	63,4115	56,1198
7	JRip	58,2031	50,3906
8	Bayes Net	53,7760	51,5625
9	Naïve Bayes	36,7188	44,6615
10	Dagging	26,0417	29,9479

Another future direction can be testing with data sets of different domains other than standard UCI repository that can be from real life data or obtained from survey on different domains.

References

[1] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-Based Syst.*, vol. 8, no. 6, pp. 373–389, 1995.

[2] P. Gray and H. J. Watson, "Present and Future Directions in Data Warehousing," *Data Base Adv. Inf. Syst.*, vol. 29, no. 3, pp. 83–90, 1998.

[3] I. Saritas, M. Koklu, and K. Tutuncu, "Performance of Classification Techniques on Parkinson's Disease," *Int. J. Adv. Sci. Eng. Technol.*, vol. 5, no. 2, pp. 9–13, 2017.

[4] M. Langaas, "Discrimination and classification. Technical report. Department of Mathematical Sciences, The Norwegian Institute of Technology. Norway," 1995.

[5] M. Koklu and K. Tutuncu, "Classification of Chronic Kidney Disease With Most Known data Mining Methods," *Int. J. Adv. Sci. Eng. Technol.*, vol. 5, no. 2, pp. 14–18, 2017.

[6] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques. Elsevier," 2011.

[7] M. Castelli, L. Trujillo, L. Vanneschi, and A. Popovic, "Prediction of energy performance of residential buildings: A genetic programming approach," *Energy Build.*, vol. 102, pp. 67–74, 2015.

[8] S. S. Gilan and B. Dilkina, “Sustainable Building Design: A Challenge at the Intersection of Machine Learning and Design Optimization,” pp. 101–106, 2015.

[9] S. A. Kalogirou, “Artificial neural networks in energy,” *Int. J. Low Carbon Technol.*, vol. 1, no. 3, pp. 201–216, 2006.

[10] A. Tsanas and A. Xifara, “Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools,” *Energy Build.*, vol. 49, no. August 2011, pp. 560–567, Jun. 2012.

[11] “Blake A.C.L. and Merz C.J. (1998). University of California at Irvine Repository of Machine Learning Databases, https://archive.ics.uci.edu/ml/datasets/Energy+efficiency, Last Access: 20.01.2017.”

[12] R. Arora and Suman, “Comparative Analysis of Classification Algorithms on Different Datasets using WEKA,” *Int. J. Comput. Appl.*, vol. 54, no. 13, pp. 21–25, 2012.

[13] K. Tutuncu and M. Koklu, “Comparison of Classification Techniques on Dermatological Dataset,” *Int. J. Biomed. Sci. Bioinforma.*, vol. 3, no. 1, pp. 10–13, 2016.

[14] J. Abellán and J. G. Castellano, “A comparative study on base classifiers in ensemble methods for credit scoring,” *Expert Syst. Appl.*, vol. 73, pp. 1–10, 2017.

[15] E. Bauer, R. Kohavi, P. Chan, S. Stolfo, and D. Wolpert, “An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants,” *Mach. Learn.*, vol. 36, no. August, pp. 105–139, 1999.

[16] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[17] Z. Barutcuoglu, “A Comparison of Model Aggregation Methods for Regression.,” *Icann*, pp. 76–83, 2003.

[18] M. Koklu and K. Tutuncu, “Performance of Classification Techniques on Medical Datasets,” *Int. J. Biomed. Sci. Bioinforma.*, vol. 3, no. 1, pp. 5–9, 2016.

[19] P. Melville and R. J. Mooney, “Diverse ensembles for active learning,” in *Twenty-first international conference on Machine learning - ICML ’04*, 2004, p. 74.

[20] S. Satu, T. Akter, and J. Uddin, “Performance Analysis of Classifying Localization Sites of Protein using Data Mining Techniques and,” pp. 860–865, 2017.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, p. 10, Nov. 2009.

[22] G. Kaur and A. Chhabra, “Improved J48 Classification Algorithm for the Prediction of Diabetes,” *Int. J. Comput. Appl.*, vol. 98, no. 22, pp. 13–17, 2014.

[23] M. Panda, A. Abraham, and M. R. Patra, “Hybrid intelligent systems for detecting network intrusions,” *Secur. Commun. Networks*, vol. 8, no. 16, pp. 2741–2749, Nov. 2015.

[24] D. Y. Mahmood and M. A. Hussein, “Intrusion Detection System Based on K-Star Classifier and Feature Set Reduction,” *Int. Organ. Sci. Res. J. Comput. Eng. Vol*, vol. 15, no. December, pp. 107–112, 2013.

[25] F. Alam and S. Pachauri, “Comparative Study of J48 , Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA,” *Adv. Comput. Sci. Technol.*, vol. 10, no. 6, pp. 1731–1743, 2017.

[26] K. Wang, “Outcome Prediction of DOTA2 Based on Naive Bayes Classifier,” no. 1994, pp. 4–6, 2017.

[27] B. Trivedi and N. Kapadia, “Modified Stacked Generalization with Sequential learning,” *Int. J. Comput. Appl.*, pp. 38–43, 2012.

[28] W. W. Cohen, “Fast effective rule induction,” *Proc. Twelfth Int. Conf. Mach. Learn.*, vol. 95, pp. 115–123, 1995.