An Experimental Study of K* Algorithm

Dayana C. Tejera Hernández
University of the Informatics Sciences/Department of Software Engineering and Management, La Habana, 10800, Cuba
Email: dtejera@uci.cu

Abstract—Machine Learning techniques are taking place in all areas of our lives, to help us to make decisions. There is a large number of algorithms available for multiple purposes and appropriate for specific data types. That is why it is required to pay special attention to decide which is the recommended technique, to use in each case. K Star is an instance-based learner that tries to improve its performance for dealing with missing values, smoothness problems and both real and symbolic valued attributes; but it is not known much information about how the way it faces attribute and class noisy, and with mixed values of the attributes in the datasets. In this paper we made six experiments with Weka, to compare K Star and other important algorithms: Naïve Bayes, C4.5, Support Vector Machines and k-Nearest Neighbors, taking into account its performance classifying datasets with those features. As a result, K Star demonstrated to be the best of them in dealing with noisy attributes and with imbalanced attributes.

Index Terms—Machine Learning techniques, K Star, k-Nearest Neighbors, Naïve Bayes, C4.5, Support Vector Machines, machine learning algorithms comparison.

I. INTRODUCTION

The amount of raw data that we store in all the spaces of our lives, is increasingly, and so, the need of understand such data for helping to the decision making. That is why Machine Learning (ML) techniques are taking place in all areas of our life. Alpaydin [1] consider that ML is optimizing a performance criterion on computers by programming them, using for the experience or example data. ML provides the technical basis of data mining (DM), which is the extraction of implicit, previously unknown, and potentially useful information from data; by means of discovering patterns that could be generalized to make accurate predictions on future data [2].

There are a lot of algorithms for DM, and they are organized according different kinds of classifications. Some of them depend on the type of representation they use for explaining its procedures or its results (rule-based, decision trees, etc.), others on the goal they have to fulfill (classification, clustering) and others depending on the moment they realize the major effort for learning. Lazy algorithms are in the last case (nonparametric approach). They defer the real work as long as possible, while others produce generalizations since they "meet" the data[2]. Instance-based (IB) learners, also called Memory-based ones, are lazy algorithms that store the training instances in a lookup table and interpolate from these [1].

IB learners are able to learn quickly from a very small dataset; while others like rule induction methods require a reasonable representation of each rule before they can be induced. An instance-based learner can begin to make useful predictions from as little as one example per class. Classification performance often exceeds 75% of the maximum possible after accepting only 25% of a complete data set. Also, is important to note that they can use continue valued features and predict numeric valued classes because they retain each instance as a separate concept, whereas other methods should partition the class into a small number of concepts. [3]

Nearest Neighbors rules are founders of this approach, followed by k–Nearest Neighbor method (k-NN), establishing the consistency of the method as k varies from one to infinity. There has been emerged some other methods following this approach, but in this paper we will be discussing the method K Star (K*), developed in 1995 by John G. Cleary and Leonard E. Trigg [4], which has had very well results treating missing values, smoothness problems and dealing with mixed values. This paper aims to perform an experimental study of this method, in order to proof its performance in relation with other important methods.

II. K* ALGORITHM

In IB classification problems, "each new instance is compared with existing ones using a distance metric, and the closest existing instance is used to assign the class to the new one" [2]. The principal difference of K* against other IB algorithms is the use of the entropy concept for defining its distance metric, which is calculated by mean of the complexity of transforming an instance into another; so it is taken into account the probability of this transformation occurs in a "random walk away" manner. The classification with K* is made by summing the probabilities from the new instance to all of the members of a category. This must be done with the rest of the categories, to finally select that with the highest probability [4].

To treat the missing values in datasets, Cleary & Trigg [4] assumed that the probability of transforming to that kind of values, is the mean of the probability of transforming to each of the specified attributes in the dataset. So, it is considered the expected distance to a random instance of that attribute.

Many authors of this area have used K* for different

classification problems [5] [6] and the results have been good.

However, there is not much information about how K* faces attribute and class noisy, and with mixed values of the attributes in the datasets. In the following section, we present an experimental study for comparing K* with some of the most influential DM algorithms, according to [7]: C4.5, Support Vector Machine (SVM), k-NN and Naïve Bayes.

II. Previous Works

Comparing ML algorithms is one of the most common methods for selecting the most appropriate one for a given situation. Many researchers has used this method evaluating their performance in one or more indicators. In this section we will be discussing some of these experiences.

The study of Kalapanidas, et al. [8] was centered on the noise sensitivity of 11 ML algorithms (0-Rule, K-NN, Linear Regression, M5, K*, MLP, Decision tables, Hyper pipes, C4.5 Decision trees, C4.5 Rules and Voting Feature Interval). They were based on two prediction problems and one classification problem. They also used four artificial datasets, the first of them implements a multivariate problem, another one a linear function, while the third and the fourth ones, refer to a non linear function; but all of them had numerical data. Five-fold cross validation experiments were carried out for each dataset. In this case, K* results were not best but neither the worst for any experiment, but is important to note that it is tested here, as a regression algorithm. Even when this study can be considered interesting, they were expressed in graphics, which curves are difficult to be differentiated, so its use could be hard.

The main goal of Er in [6] was to propose a method for accurate prediction of at-risk students in an online course, taking into account log data of the LMS used. All attributes had continuous values (0 - 100), except for one that could be *yes or no*. Also, they had a common characteristic: were time-varying. Were evaluated three ML algorithms: Naïves Bayes, K* and C4.5. The best rates of accuracy, sensitivity and precision, of the ML algorithms, were reached for K*. Anyway, this results can only be taken into account with this type of datasets; and just for discerning between this algorithms, according to their performance in this indicators.

Vijayarani and Muthulakshmi [9] compared Lazy (the basic Instance Based Learner (IBL), k-NN and K*) and Bayesian (Bayesian Net and Naïve Bayes) classifiers for text mining problems. They used a dataset of 80000 instances and four attributes. Even when lazy classifiers performed better, K* got the highest error rates, and resulted the worst in the most of the indicators of accuracy (% of correctly/incorrectly classified instances, TP rate, ROC Area and Kappa Statistics). This study offers interesting values of K*, but are not declared details of the kind of values that the dataset used contains.

Douglas et al. in [10] evaluated six machine learning algorithms (K*, Naïve Bayes, Suport vector classifiers, Decision tree, AdaBoost classification, and Random forest) over a range of complexity; for identifying which is the most accurate in mining functional neuroimaging data. K* accuracy was of 86% (the secondly worst value after Support Vector Machine).

These studies shows us a heterogeneous set of results of the performance of K*: sometimes are not relevant (nor the best neither the worst), others is unquestionable the best and in others definitively the worst. These differences are due to the heterogeneous composition of the datasets evaluated in each of these studies; but it is difficult to generalize, because most of them are focused on very specific domain; sometimes the kind of data are not detailed; several indicators are evaluated with the same dataset; none of them refers to the use of a statistical analysis for evaluating the relevance of the results.

The experiments developed in this research compares K* with other four ML algorithms, with the intention of suggesting its use for specific data types in a wider level of generalization, evaluating the relevance the obtained results and offering sufficient information for other analysis and conclusions.

IV. Experimental Study

To facilitate the reference to each algorithm during the experiments, we have defined identifiers for all of them (See Table 1).

Table 1 Identifiers of the algorithms.

Identifier	Algorithm
1	Naïves Bayes
2	K–Nearest Neighbors (IBk)
3	K*
4	C4.5 (J48)
5	Lib SVM

Weka [11] was the data mining tool utilized, and its parameters were left as default. The datasets were taken from the Knowledge Extraction based on Evolutionary Learning (Keel) repository [12]. The type of experiment realized was 10-Fold Cross-Validation with ten iterations.

Each experiment was focused on evaluating a type of data.

A. Standard Examples

At the beginning a test with nineteen datasets of standard values were performed (See Table 2). Every one defines a supervised classification problem [12].

Table 2 Description of the datasets used for standard attribute values.

Dataset	Number of examples	Number of classes	Number of attributes (Real/Integer/ Nominal)	Missing values
Appendicitis	106	2	(7/0/0)	No
Australian	690	2	(3/5/6)	No
Automobile	150	6	(15/0/10)	Yes
Balance	625	3	(4/0/0)	No
Bands	365	2	(13/6/0)	Yes
Breast	277	2	(0/0/9)	Yes
Bupa	345	2	(1/5/0)	No
Car	1728	4	(0/0/6)	No
Cleveland	297	5	(13/0/0)	Yes
Contraceptive	1473	3	(0/9/0)	No
Crx	653	2	(3/3/9)	Yes
Dermatology	358	6	(0/34/0)	Yes
Ecoli	336	8	(7/0/0)	No
Mammographic	830	2	(0/5/0)	Yes
Monk-2	432	2	(0/6/0)	No
Post-operative	87	3	(0/0/8)	Yes
Saheart	462	2	(5/3/1)	No
Tae	151	3	(0/5/0)	No
Wisconsin	683	2	(0/9/0)	Yes

However, how the composition of the datasets is so heterogeneous, is too difficult to make a decision about the effectiveness of K* for the different aspects we are interested. That is why we present below, 5 new experiments, each one for a specific feature.

B. Missing values

For comparing the efficacy of these algorithms treating the attributes with missing values, we have performed an experiment with thirteen datasets, seven with originally missing values and six induced, that is the 10% of the values of standard datasets were randomly removed. In both cases, a selection criterion was avoiding those datasets which had mixed values (See Table 3).

Table 3 Description of the datasets with missing values.

Dataset	Number of examples	Number of classes	Number of attributes (Real/ Integer/ Nominal)	% Missing values (Examples)	Induced
Post-operative	90	3	(0/0/8)	3.33	No
Breast	286	2	(0/0/9)	3.15	No
Cleveland	303	9	(13/0/0)	1.98	No
Dermatology	366	6	(0/34/0)	2.19	No
Housevotes	435	2	(0/0/16)	46.67	No
Mammographic	961	2	(0/5/0)	13.63	No
Wisconsin	699	2	(0/9/0)	2.29	No
Ecoli+MV	336	8	(7/0/0)	48.21	Yes
Iris+MV	150	3	(4/0/0)	32.67	Yes
Magic+MV	1902	2	(10/0/0)	58.20	Yes
Pima+MV	768	2	(8/0/0)	50.65	Yes
Wine+MV	178	3	(13/0/0)	70.22	Yes
Shuttle+MV	2175	7	(0/9/0)	55.95	Yes

Two previous tests were done, one for original datasets and the other for induced ones.

C. Noise

For analyzing how these algorithms carry out the noise treatment, two experiments were realized: one for datasets with noisy attributes and the second for noisy classes.

The first experiment was done with a group of nine datasets composed of four ones with a 20% of noise and other five only with a 15 %.

Table 4 Description of the datasets with noisy attributes.

Dataset	Number of examples	Number of classes	Number of attributes (Real/ Integer/ Nominal)	% Noisy Attributes
Contraceptive	1473	3	(0/9/0)	15
Glass	214	7	(9/0/0)	15
Iris	150	3	(4/0/0)	15
Wdbc	569	2	(30/0/0)	15
Wine	178	3	(13/0/0)	20
Ecoli	336	8	(7/0/0)	20
Pima	768	2	(8/0/0)	20
Sonar	208	2	(60/0/0)	20
Yeast	1484	10	(8/0/0)	20

The noisy class experiment was carried out with nine datasets with 20% of noise, avoiding those with mixed values.

Table 5 Description of the datasets used for the experiment of the datasets with noisy classes.

Dataset	Number of examples	Number of classes	Number of attributes (Real/ Integer/ Nominal)
Contraceptive	1473	3	(0/9/0)
Ecoli	336	8	(7/0/0)
Glass	214	7	(9/0/0)
Iris	150	3	(4/0/0)
Pima	768	2	(8/0/0)
Sonar	208	2	(60/0/0)
WDBC	569	2	(30/0/0)
Wine	178	3	(13/0/0)
Yeast	1484	10	(8/0/0)

D. Imbalanced attributes

This experiment was have performed with thirteen datasets of different levels of imbalance ratio.

Table 6 Description of the datasets used for the experiment of the datasets with imbalanced attributes.

Dataset	Number of examples	Number of attributes (Real/ Integer/ Nominal)	Imbalance Ratio
Cleveland-0_vs_4	177	(13/0/0)	12.62
Ecoli-0_vs_1	220	(7/0/0)	1.86
Ecoli	336	(7/0/0)	71.5
Ecoli1	336	(7/0/0)	3.36
Glass-0-1-5_vs_2	172	(9/0/0)	9.12
Glass	214	(9/0/0)	8.44
Glass1	214	(9/0/0)	1.82
Haberman	306	(0/3/0)	278
Hayes-Roth	132	(0/4/0)	1.7
Led7digit-0-2-4-5-6-7-8-9_vs_1	443	(7/0/0)	10.93
Pageblocks	548	(10/0/0)	164
Wine	178	(13/0/0)	1.5
Yeast-0-3-5-9_vs_7-8	506	(8/0/0)	9.12

E. Mixed values

This experiment was done with twelve datasets with mixed values, but has a limitation, three of them contain missing values.

Table 7 Description of the datasets used for the experiment of the datasets with mixed values attributes.

Dataset	Number of examples	Number of classes	Number of attributes (Real/ Integer/ Nominal)	Missing values
Abalone	4174	10	(7/0/1)	No
Australian	690	2	(3/5/6)	No
Automobile	150	6	(15/0/10)	Yes
Crx	653	2	(3/3/9)	Yes
German	1000	2	(0/7/13)	No
Heart	270	2	(1/12/0)	No
Hepatitis	80	2	(2/17/0)	Yes
Ionosphere	351	2	(32/1/0)	No
Lymphography	148	4	(0/3/15)	No
Page-blocks	5472	5	(4/6/0)	No
Saheart	462	2	(5/3/1)	No
Vowel	990	11	(10/3/0)	No

For each experiment the Friedman's test [13] was performed, and as a result the five algorithms were increasingly arranged, taking into account that the smaller value should be associated to the algorithm with the best performance. Also, a P-value has been obtained, which is used for determining if the differences between the algorithms are statistically significant for the set value of α = 0.05. If the P-value > α, then is rejected the hypothesis, so the difference between the algorithms is not statistically significant.

If the opposite happens, then it must be analyzed the result of Holm's procedure, which extracts the best algorithm from the ranking, to be compared with the others (as a control classifier); for analyzing the significance of that difference more specifically.

IV. RESULTS AND DISCUSSION

In the experiment performed for standard examples, C4.5 (4) was the algorithm with the best performance according to the ranking. K* (3) is in the fourth place (See Table 8). The P-value obtained was 0.2851441771134994, which implies that the null hypothesis was rejected; so we can conclude that is a good performance anyway, because their differences are not significant.

Table 8 Average rankings of the algorithms, for the experiment of he standard datasets

Algorithm	Ranking
3	3.1578947368421053
1	2.7105263157894743
2	3.157894736842105
4	2.4736842105263155
5	3.5

C4.5 (4) was the algorithm with the best performance in the missing values experiments too; but K* (3) ascended to the third place (See Table 9). This time the P-value = 0.3659628532452208, so their difference are neither statistically significant.

Table 9 Average rankings of the algorithms, for the experiment of the datasets with missing values.

Algorithm	Ranking
3	3.00000000000000004
1	2.76923076923077
2	3.076923076923077
4	2.4615384615384617
5	3.692307692307692

As described in the previous section, two new tests were done, one for original datasets and the other for induced ones. The performance of K* was better in the first group of datasets than in the second. Note that the induced datasets have a major percent of missing values than the originals (See Table 3), so it could be supposed that while the concentration of missing values in the data be minor, better should be the performance of K*. But it must be demonstrated in further works.

The analysis of the remaining experiments results will continue with noisy ones, where surprisingly, K* is in the first place (See Table 10), in contrast with the ones that were obtained in Kalapanidas [8] work. And also, this time the differences are significant, because Friedman´s test threw a P-value < α (0.014334880309170184). Taking into account this result, is needed to make a deeper analysis in respect with how significant is the difference between K* and the rest algorithms, according to the Holm's test.

Table 10 Average rankings of the algorithms, for the second experiment of datasets with noisy attributes.

Algorithm	Ranking
3	1.888888888888888888
1	3.1111111111111111116
2	4.0
4	2.222222222222222223
5	3.777777777777777777

Holm's test rejects those hypotheses which p-value ≤ 0.025. If we observe Table 11, we will discover that the p-value obtained for the algorithms 2 and 5 fits this condition, so it is possible to affirm that the differences between K* and they are statistically significant.

Therefore, we have checked Shaffer's procedure (See Table 12), which rejects the hypotheses that have a p-value ≤ 0.005. In this case, only the second algorithm fits that condition, so we can finally ensure that K* is statistically significant better than k-NN for noisy attributes.

Table 11 Holm / Hochberg Table for α = 0.05.

i	algorithm	$z = \dfrac{R_0 - R_i}{SE}$	p	Holm/ Hochberg/ Hommel
4	[2]	2.8323527714997366	0.004620683963359944	0.0125
3	[5]	2.5342103744997617	0.011270104849283176	0.0166666
2	[1]	1.6397831834998464	0.10105025592540978	0.025
1	[4]	0.44721359549995815	0.6547208460185769	0.05

Table 12 Holm / Shaffer for α = 0.05.

i	algorithm	$z = \dfrac{R_0 - R_i}{SE}$	p	Holm	Shaffer
10	[3] vs. [2]	2.832352771499736	0.004620683963359944	0.005	0.005
9	[3] vs. [5]	2.5342103744997617	0.011270104849283176	0.0055556	0.0083333
8	[2] vs. [4]	2.3851391759997758	0.01707266109378837	0.00625	0.0083333
7	[4] vs. [5]	2.0869967789998034	0.03688842570704993	0.0071428571435	0.0083333
6	[3] vs. [1]	1.6397831834998464	0.10105025592540978	0.0083333	0.0083333
5	[1] vs. [4]	1.1925695879998883	0.23303798227390524	0.01	0.01
4	[1] vs. [2]	1.1925695879998872	0.23303798227390565	0.0125	0.0125
3	[1] vs. [5]	0.8944271909999151	0.371093369522698	0.0166666	0.0166666
2	[3] vs. [4]	0.44721359549995815	0.6547208460185769	0.025	0.025
1	[2] vs. [5]	0.2981423969999721	0.765594483995764	0.05	0.05

Otherwise, in noisy class experiment K* descended to the third place again (See Table 13), and C4.5 is the algorithm with the best performance. The P-value > α (0.05998131067845795); so the differences are not significant.

Table 13 Average rankings of the algorithms, for the experiment of datasets with noisy classes.

Algorithm	Ranking
1	2.44444444444444446
2	2.11111111111111111
3	3.05555555555555554
4	3.22222222222222228
5	4.1666666666666666

For imbalanced data, K* reaches the first place again, followed by C4.5 (4). But how the p-value < α (0.49563823253249994), then we can not ensure that its differences are statistically significant.

Table 14 Average rankings of the algorithms, for the experiment of datasets with imbalanced attributes.

Algorithm	Ranking
3	2.384615384615385
1	3.0769230769230766
2	3.3076923076923075
4	2.8461538461538467
5	3.384615384615384

And in the mixed values experiment, K* fell back into the fourth place of the ranking list, but we can neither we can neither conclude that it is not so different to the others (0.48296481840196503 = p-value < α)

Table 15 Average rankings of the algorithms, for the experiment of datasets with mixed values attributes

Algorithm	Ranking
3	2.9166666666666666665
1	2.8333333333333333333
2	2.75
4	2.75
5	3.750000000000000004

In the Table 16 a summary of the principal results are showed. Even when it reflects that K* wins two times, it can only be ensured that it performs better than the other evaluated learners, with noisy attributes. Anyway its performance is also good on datasets with imbalanced attributes, class noise, and even with the rest others, because it is demonstrated that theirs differences were not significant. Anyway, for datasets containing mixed values and standard attributes, k-NN and C4.5 respectively are much better.

Table 16 Summary of the experiments results

Algorithms	Standard Experiment	Missing values	Attribute noise	Class noise	Imbalanced attributes	Mixed values
Naïve Bayes	2	2	3	2	3	3
k-NN	3	4	5	1	4	1
K*	4	3	1	3	1	4
C4.5	1	1	2	4	2	1
(SVM	5	5	4	5	5	5

V. CONCLUSIONS

An experimental study of K* algorithm is reported in this paper, in order to prove its performance in relation with other important methods (Naïve Bayes, C4.5, Support Vector Machines and k-Nearest Neighbors).

The examination of four previous experiences in comparing K* efficiency with the ones of other learners, showed a heterogeneous set of results due to the diverse composition of the datasets evaluated in each of these studies. Taking this into account the objective of this paper were centered on suggesting the use of K* for specific data types in a wider level of generalization, evaluating the relevance the obtained results and offering sufficient information for other analysis and conclusions.

The experiments type was 10-Fold Cross-Validation with ten iterations, were carried out on Weka, with five datasets taken from Keel repository. Each of them was focused on evaluating a type of data (standard examples, noisy attributes and classes, missing, imbalanced and mixed values).

For attribute noise and imbalanced attributes datasets, K* reached the first place in the ranking of performances; the third for the missing values and noisy classes datasets, and the fourth place for standard examples and mixed values datasets. The use of Friedman's Test and Holm's and Shaffer's procedures allowed us to demonstrate how, of all the previous mentioned results, is statistically significant only the superiority of K* over k-NN for attribute noise. So, even in the cases in which K* got the last places, it can be considered its use for all of this data types, because their differences with the other's performances are not statistically relevant.

REFERENCES

[1] Alpaydn, E., *Introduction to Machine Learning*, T. Dietterich, Editor. 2010: London, England.

[2] Witten, I.H., E. Frank, and M.A. Hall, Data Mining. Practical Machine Learning tools and techniques, ElSevier, Editor. 2011.

[3] Martin, B., Instance-Based Learning: Nearest Neighbour with Generalisation, in Department of Computer Science. 1995, University of Waikato: Hamilton, New Zeland. p. 83.

[4] Cleary, J. and L. Trigg, K*: An Instance-based Learner Using an Entropic Distance Measure, in 12th International Conference on Machine Learning. 1995. p. 108-114.

[5] Uzun, Y. And G. Tezel, Rule Learning With Machine Learning Algorithms And Artificial Neural Networks. Journal of Seljuk University Natural and Applied Science, 2012. 1(2).

[6] Er, E., Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100. International Journal of Machine Learning and Computing, 2012. 2(4): p. 279.

[7] Wu, X., et al., Top 10 algorithms in data mining. Springer, 2008. 14: p. 1-37.

[8] Kalapanidas, E., et al., Machine Learning Algorithms: a Study on Noise Sensitivity. First Balkan Conference in Informatics, 2003 November; pp. 356-365.

[9] Vijayarani, S.and Muthulakshmi, M., Comparative Analysis of Bayes and Lazy Classification Algorithms. International Journal of Advanced Research in Computer and Communication Engineering, 2013 August; 2(8): 3118-3124.

[10] Douglas, P. K., et al., Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief. Neuroimage. 2011 May 15; 56(2): 544-553.

[11] Hall, M.A., et al., WEKA. 2009.

[12] Alcalá-Fdez, J., et al., KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Multiple-Valued Logic & Soft Computing, 2011. 17: p. 255-287.

[13] Dêmsar, J., Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research, 2006. 7: p. 1-30.

Author's Profile

Dayana C. Tejera Hernández was born in La Habana, Cuba, on 1984 August 9th.

Educational degree:

• Informatics Sciences Engineer, 2007, University of the Informatics Sciences, Havana, Cuba.

• Master on Educational Technology, e-Learning and Knowledge Management, 2012, University of Balearic Islands, Palma de Mallorca, Spain. The author's major field of study is now Educational Data Mining for Informatics Talent Management.

She works at the University of Informatics Sciences, as Methodologist and the Principal Professor of the courses Software Engineering I and II. She also has been the Leader of the Department of Software Engineering and Management, Principal Professor of the Academical 5th Year of the career, and professor of the grade courses: Scientific Research Methodology, Software Management, Software Requirement Engineering and also of the post-grade course: Fundamentals of Agile Software Development. She has presented works on conferences like XI International Conference of Educational Sciences (CECEDU), V Iberoamerican Virtual of Quality on Virtual and Distance Education (Educ@), International Conference on Open Source Systems (OSS 2014, ,) Conference International Convention and Feria Informatics, UCIENCIA, etc.