Analyzing the Discretization and Feature Selection Problems applied to Petroleum Data Sets

Luis Carlos Molina
Mexican Institute of Petroleum
PIMAyC – PIYYAC
Eje Central 152
07730 Mexico, D.F.
Email: lcmolina@imp.mx

Lluís Belanche
Technical University of Catalonia
Software Department
Jordi Girona 1-3 C6 214
08004 Barcelona, Spain.
Email: belanche@lsi.upc.es

***Abstract*—In petroleum databases one usually has to deal with problems like noise, high variance, unbalanced classes, missing values, small or large number of cases, ambiguity or inconsistencies, irrelevant and/or redundant features, etc. In consequence, when the data are analyzed by means of data mining, machine learning or soft computing techniques oriented to classification, it is often the case that we must confront a discretization or a feature selection process, or both. In this work we analyze two efficient algorithms, one to discretize and the other to select a subset of relevant features. The same criterion in the induction process is used, both to discretize and to select the relevant features, and these two steps are performed in this particular order. The results obtained (as measured by means of C4.5 error) suggest that this may be a good way of tackling these important problems.**

***Index Terms*—Petroleum Data Sets, Feature Selection, Discretization.**

I. Introduction

In the last decades, new petroleum processes have generated large quantity of data with the objective of capturing specific and more detailed behaviors. In this vein, methods as data mining, machine learning or soft computing are joining their efforts together to traditional techniques (e.g. statistics) to analyze this big quantity of data [9]. Petroleum datasets are themselves very interesting to analyze due to the large quantity of problems that are present: interpolation errors, noise, constant human manipulation, high variance, unbalanced classes, uncertainty, vagueness, missing values, small number of cases, ambiguity, irrelevant and/or redundant data, as well as characteristics related to economic, political and environmental aspects that are also important to be considered.

When the data are analyzed by means of data mining, machine learning or soft computing techniques oriented to classification, it is often the case that we must confront a discretization or a feature selection process, or both. In the real world most of the datasets are composed by continuous data with some irrelevant and/or redundant features. Many of the classification algorithms (inducers) require discretized data to work, also is recommended to do some feature selection process to take some advantages [4]. One of the most important reasons to use soft computing or symbolic techniques in the petroleum industry is economic. These techniques can be used to simulate or understand chemical or geophysical processes, obtain knowledge, explanations of cases, simulations, pattern recognition analysis, etc. Certainly the economical cost can be reduced and the results can improve the petroleum processes. In the petroleum industry, we can observe several strategic activities: Exploration, Drilling, Exploitation, Refining, Petrochemical Transformation, Storage, and Transportation [10]. Each stage is composed of several and complex processes. When we talk about petroleum datasets, we refer to the data acquisition generated in some part of a petroleum process. In this work only examples in exploration, drilling and exploitation were used due to the facility of the information access and the available expert knowledge.

Our proposal is focused on datasets where supervised paradigms are applied. The main contribution is to combine two effective algorithms, one to discretize and other to select relevant features. In both cases, the criterion used by a classifier in the induction process is also used to discretize and to select relevant features. We successfully tried to improve the performance, at least in the petroleum domain, over existing discretization and feature selection algorithms. This work examines the main components that are considered in these processes, and is organized as follows: in the next two sections, we briefly review the discretization and feature selection processes. Next, we describe the used petroleum datasets and the experimental set up. Finally, we describe the experiments performed and discuss the results.

II. Discretization

***Discretization* is the process of converting continuously valued data into discrete data by assigning ranges of real values (established by *cut–points*) to an ordered set of discrete labels. The choice of these intervals is a critical issue as too many intervals impair the comprehensibility of the models and too few hide important features of the variable distribution. Most often the user must specify the number of intervals, or provide some heuristic rule to be used. The discretization of the target variable values provides a different granularity of predictions that can be considered more comprehensible. Discretization should significantly reduce the number of possible values of the continuous attribute since large number of possible attribute values contributes to slow and ineffective process of inductive machine learning [3].**

Discretization algorithms can be divided into two categories:

- ***Unsupervised algorithms* that discretize attributes without taking into account respective class labels. The representative algorithms are equal–width and equal–frequency discretizations.**
- ***Supervised algorithms* discretize attributes by taking into account the class-attribute interdependence.**

The method proposed in this work concerns automatically finding the optimal number and width of these intervals by means of Evolutionary Strategies [1]. The basic idea is to generate populations of cut points and intervals that are evaluated by a fitness function, which uses the same classification algorithm. In the present case, the fitness function selected is the error of the C4.5 criterion. This criterion is used in the C4.5 algorithm for building decision trees, and it is based on the notion of *information gain* [12].

III. Feature Selection

The most important feature selection problem in terms of supervised inductive learning is: given a set of candidate features select a subset defined by one of three approaches: a) the subset with a specified size that optimizes an evaluation measure, b) the subset of smaller size that satisfies a certain restriction on the evaluation measure and c) the subset with the best commitment among its size and the value of its evaluation measure (general case). The generic purpose pursued is the improvement of the inductive learner, either in terms of learning speed, generalization capacity or simplicity of the representation. It is then possible to understand better the results obtained by the inducer, diminish its capability of storage, reduce the noise generated by irrelevant or redundant features and eliminate useless knowledge.

A Feature Selection Algorithm (FSA) is a computational solution that is motivated by a certain definition of *relevance*. However, the relevance of a feature –as seen from the inductive learning perspective– may have several definitions depending on the objective that is looked for. An irrelevant feature is not useful for induction, but not all relevant features are necessarily useful for induction [2].

The FSAs can be classified according to the kind of output they yield:

- Those algorithms giving a (*weighed*) linear order of features.
- Those algorithms giving a *subset* of the original features.

Both types can be seen in an unified way by noting that in *subset* the weighting is binary.

In the present paper, the proposal presented is centered in FSAs tackling the feature selection problem of type *subset*. We use a Sequential Floating Feature Selection (SFFS) [11], with a modified stopping criterion (see below).

IV. Petroleum Datasets Used

The experiments are composed by eight different petroleum datasets in order to analyze the role of discretization and feature selection. Representative datasets of each process were selected to show some common problems in these analysis. For all datasets, the names of the attributes have been changed to protect the information, however, an explanation of each problem domain will be given. The datasets structure is commonly well–known by the experts. See Table I for numerical details.

A. Pollution (POLL)

Some times, near the well's neighborhood, a determinate area is established to deposit pollution waste. The objective of this process is to detect different pollution chemical elements in a specific zone in various levels. The zone is delimited by thirteen sample points that detect pollutants to four levels of ground. The zone was sampled during three years in intervals of six months each. The class attribute shows the pollution rate and is called Total Petroleum Hydrocarbons (TPH). It is here divided in low, low–medium, high–medium and high pollution.

B. Remediation (RMD)

An experiment was made to determine contamination levels using remediation method of placing straw and bacteria as main component. The objective is to determine up to where the bacteria influence the hydrocarbon degradation. Four bathtubs with polluting agents were covered by blankets to observe their levels of degradation of contamination through time. The bathtubs were sampled during six months in intervals of one month. The class–attribute shows the pollution rate and is called Total Petroleum Hydrocarbons (TPH). It was divided in low–medium, high–medium and high pollution.

C. Well Logs (WL)

Petrophysical properties can be obtained from logging instruments that are lowered in the wells and by core analysis on reservoir rock material that is obtained from the well with a hollow drill bit. Permeability is a critical petrophysical variable for both petroleum geology and petroleum engineering. The experiment is composed by six well logs obtained from Smackover Formation in Big Escambia Creek field, Alabama, USA. Only the porosity and spatial information were used to describe permeability. The class was divided in poor, medium–low, medium–high, and high.

D. Lithofacies (LF)

In the reservoir characterization and reservoir simulation, the facies properties that are most important are the petrophysical characteristics that control the fluid behavior in the facies. The lithofacie resumes the main mineralogic properties (texture, mineralogy, grain size, and the depositional environment that produced it) of the rock. The lithofacies are facies characterized by the distribution of mineral grains and sedimentary rock types. Several models have been used to estimate lithofacies from geophysical data, well logs and other defined curves. In this domain 4 datasets were considered. The lithofacies (class label) were determined by experts.

E. Pressure–Vapor-Temperature (PVT)

The term PVT stands for the relation between Pressure–Vapor–Temperature. This dataset is constituted by seventeen chemical compounds formed mainly by H, He, B, C, N, O, Ne, S, and Cl. One record consists of the number of molecules of each chemical element that composed the compound, two constants a and b, molecular mass, boil temperature. The classes were divided in low, low–medium, high–medium and high critical temperature.

TABLE I

Characteristics of petroleum datasets used in the experiments (*cont/nom* = number of continuous/nominal features, *Maj. Class* = Majority Class).

Dataset	Attribute Type		Class	Missing values	Cases	Maj. Class
	cont	nom				
POLL	61	4	4	49.3%	192	52.5%
RMD	11	3	3	11.9%	280	40.0%
WL	4	2	4	0.0%	981	48.9%
LF-2	22	1	10	0.0%	3050	52.6%
LF-5	25	1	5	0.0%	2335	50.6%
LF-23	21	1	4	0.0%	1282	49.9%
LF-25	25	1	7	0.0%	1931	55.6%
PVT	4	18	4	0.0%	355	34.0%

V. Experimental Setup

Due to confidentiality of the information, some modifications were done to the datasets. The attribute names were changed to $A_1, \ldots, A_n$ and the class names to $0, 1, 2, \ldots, n$, where 0 means the lower value and *n* the higher value. The missing values were substituted by "?", in order to be treated as any other value (see below).

An important decision is that there exist two ways of setting up the involved experiments: 1) first discretize and then select relevant attributes or, 2) first select relevant attributes and then discretize them. We choose the first case, because it has some advantages: there is an evident loss of information when the discretization is made (e.g. originally different values are now consider as equal). When a posterior feature selection process is realized, this loss of information is then taken into account. Should we first select features, the posterior discretization process could alter the relations between attributes. On the other hand, some feature selection algorithms that work effectively on discrete data can be used [8].

Our proposal considers the use of two tools: an evolution strategies discretization algorithm [13], using C4.5 error as fitness function (ES–C4.5) and the sequential floating feature selection (SFFS–C4.5) [11] (also using C4.5 error as the criterion to select subsets of relevant features).

In each selection step, SFFS performs a forward step followed by a variable number (possibly null) of backward ones. In essence, a feature is first unconditionally added and then features are removed as long as the generated subsets are the best among their respective size. The algorithm is so-called because it has the characteristic of *floating* around a potentially good solution of the specified size. The original stopping condition of the algorithm needs the setting of a desired number of features. Since this number is unknown a priori, we modified it so that the algorithm stops whenever a number of consecutive floating steps do not yield an improvement. For the present experiments, this number is set to three, based on preliminary experimentation.

VI. Experiments and Discussion

To realize the evaluations, we use the *J48* algorithm, a java version of the original C4.5 release 48 by means of the the Weka tool [14].

A. Experiments on discretization

To establish the same criterions to evaluate the results, the first step was discretize the whole dataset and afterwards, we evaluated, by means of C4.5 error, the discretized datasets using 5–fold cross–validation and using training set as test set.

We discretize in four forms: i) using the discretization made for the C4.5 algorithm, ii) simultaneous discretization using evolution strategies (ES–C4.5), iii) MDLP [5] and, iv) ChiMerge [6].

ChiMerge method is based on the statistical χ^2 approach for supervised discretization. The algorithm begins by placing each numeric value into its own class and merge them according to a χ^2 test applied to neighboring classes. The hypothesis tested is that two adjacent classes are independent, which is based on the comparison between the expected and observed frequencies of values found in the corresponding classes. The merging procedure is applied until a χ^2-threshold is reached. The significance level was set to 0.9.

The Minimum Description Length Principle (MDLP) is based on a recursive entropy minimization heuristic for controlling the generation of decision trees. For evaluating each cut point, the data are discretized in two intervals and the resulting class information entropy is calculated. A coding scheme is defined which enables the comparison of information gains obtained with different cut points of the studied attribute, in terms of their codified lengths. Then, they are accepted or rejected according to the MDLP criterion.

For the missing values treatment, in MDLP and ChiMerge case, if a numerical attribute has missing values, an additional category was created for them, and the above discretization procedure is applied just to the instances for which the attribute's value is defined, in other words, the calculation would simply omit this attribute [14].

For all experiments, the ES–C4.5's parameters were the same. Based in previous experiments, we use the next value parameters to evolution strategies algorithm: Initial σ = (0.001,0.1), Total number of generations = 40, Desired Fitness = 0, Initial children μ = 150, Initial parents λ = 350, Criterion to generate populations = (μ,λ), Criterion to initialize the generations = λ–*Random*. For more detail of these parameters see [13]

The results obtained can be seen in Table II. In general terms, ES–C4.5 gave satisfactory results for nearly all datasets. The most important results were obtained in Pollution (POLL), Well Logs (WL) and Pressure–Vapor–Temperature (PVT). For the case of LF-2, the only one with worse performance, in a subsequent experiment we increased the number of generations until 200. The new error was reduced from 6.54 % to 0.45% for the train set and from 15.54% to 1.11% for 5-CV.

TABLE II

Classification errors obtained with some discretization algorithms applied to C4.5 (*5–CV* = 5–fold cross–validation. *Train* = represents to use the training set as test set. *Arity* = mean number of categories per attribute).

Data Set	Disc. Method	C4.5		Arity
		5-CV	Train	
POLL	C4.5	30.16	12.29	
	MDLP	32.40	21.78	1.46
	ChiMerge	32.96	22.34	1.56
	ES-C4.5	**24.02**	**10.61**	1.36
RMD	C4.5	**11.42**	6.42	
	MDLP	14.28	12.85	3.71
	ChiMerge	14.64	7.14	7.14
	ES-C4.5	**11.42**	**5.00**	2.71
LF-2	C4.5	**1.60**	**0.06**	
	MDLP	2.19	1.40	6.56
	ChiMerge	2.26	1.37	6.65
	ES-C4.5	15.54	6.54	2.00
LF-5	C4.5	1.75	**0.04**	
	MDLP	1.54	0.68	7.80
	ChiMerge	1.71	0.94	6.50
	ES-C4.5	**1.11**	0.17	1.80
LF-23	C4.5	2.34	0.54	
	MDLP	2.65	1.40	5.00
	ChiMerge	2.34	1.01	5.90
	ES-C4.5	**1.32**	**0.31**	1.77
LF-25	C4.5	1.91	**0.20**	
	MDLP	2.22	1.24	6.92
	ChiMerge	1.60	1.08	6.50
	ES-C4.5	**0.93**	0.36	1.76
WL	C4.5	15.47	**8.24**	
	MDLP	13.74	13.23	6.25
	ChiMerge	16.80	14.56	5.5
	ES-C4.5	**10.89**	8.85	4.5
PVT	C4.5	12.11	7.60	
	MDLP	8.45	8.45	2.75
	ChiMerge	9.85	5.07	4.00
	ES-C4.5	**7.04**	**4.78**	2.50

The results thus show that ES–C4.5 is very effective, robust, and capable of outperforming classical discretization techniques in the data mining, soft computing or machine learning fields. Moreover, this increased performance is obtained with discretizations having a much smaller number of categories/attribute (arity), therefore, with much simpler models. This is crucial when the results are interpreted by human experts (humans have difficulty handling more than 7-9 categories simultaneously).

B. Experiments on Feature Selection

For the Feature Selection experiments, we tried to select a subset with the more relevant features in each data set. We use as reference the error given for ES-C4.5. Afterwards, we compare the well–known Relief algorithm [7] versus our modified version of (SFFS) [11].

Briefly, Relief chooses randomly an instance A and determines its *near hit* and its *near miss* in relation to S. The former is the closest instance to A among all the instances in the same class of

TABLE III

CLASSIFICATION ERRORS OBTAINED WITH RELIEF AND SFFS-C4.5 APPLIED TO THE C4.5 ALGORITHM (*OF* = ORIGINAL FEATURES. *FS Method* = FEATURE SELECTION METHOD USED. */FS* = FEATURES SELECTED). THE NUMBER AFTER AN SLASH IS THE ARITY USED.

Data Set	OF	FS Method	5-CV			Train		
			ChiMerge	MDLP	ES-C4.5	ChiMerge/FS	MDLP/FS	ES-C4.5/FS
POLL	65	C4.5	32.96	32.40	24.02	22.34	21.78	10.61
		Relief	33.85	32.40	24.02	24.47	21.78	10.61
		SFFS-C4.5	30.72	29.60	24.02	13.96/31	14.50/57	10.61/60
RMD	14	C4.5	13.92	14.28	8.57	7.14	12.85	3.92
		Relief	13.57	13.57	8.57	7.85	11.42	3.92
		SFFS-C4.5	13.57	12.50	8.57	7.14/4	10.35/5	3.92/8
LF-2	23	C4.5	2.26	2.19	1.11	1.37	1.40	0.45
		Relief	2.55	2.42	1.11	1.47	1.31	0.45
		SFFS-C4.5	2.22	2.81	1.11	1.37/18	1.04/18	0.45/18
LF-5	26	C4.5	1.71	1.54	1.11	0.94	0.68	0.17
		Relief	1.71	1.51	1.11	0.94	0.64	0.17
		SFFS-C4.5	1.84	1.49	1.02	0.72/21	0.68/21	0.51/21
LF-23	22	C4.5	2.34	2.65	1.32	1.01	1.40	0.31
		Relief	8.04	2.65	1.32	5.73	1.40	0.31
		SFFS-C4.5	3.27	2.26	2.65	2.80/5	0.78/22	0.31/17
LF-25	26	C4.5	1.60	2.22	0.93	1.08	1.24	0.36
		Relief	1.60	2.22	1.55	1.08	1.03	0.88
		SFFS-C4.5	1.60	1.86	0.93	0.82/21	0.93/21	0.36/21
WL	6	C4.5	16.80	13.74	10.89	14.56	13.23	8.85
		Relief	16.70	13.84	10.89	14.76	13.23	8.85
		SFFS-C4.5	16.59	13.84	10.89	14.35/4	13.23/3	8.85/5
PVT	22	C4.5	9.85	8.45	7.04	5.07	8.45	4.78
		Relief	9.85	8.45	7.04	5.07	8.45	4.78
		SFFS-C4.5	9.29	8.73	7.04	5.63/17	8.45/16	4.78/17

A. The latter is the closest instance to A among all the instances in a different class. The importance (relevance) of each feature is then incremented (decremented) proportionally to its ability to separate A from its near miss (near hit).

Due to the fact that Relief gives results in a weighed form and SFFS in binary form, we first obtained the optimal feature subset size k as established for SFFS, and then we select the first k features in the order given by Relief. The results are shown in Table III.

Most of the times, the results given by Relief and SFFS–C4.5 were quite similar. Nevertheless, for LF-5 Relief archived better outcome and for LF-25, SFFS-C4.5 algorithm obtained the best results. For the others databases the results given by both algorithms were exactly the same. This is pretty interesting, given that one of the most important problems of feature selection algorithms is to choose the number of attributes. In the same way, an important advantage of those algorithms is that they provide rankings of features which help us to identify the contributions of each ones over the dataset.

Finally, we would like to mention that during these experiments we observed that the final errors given by ES–C4.5 and SFFS–C4.5 were, most of the times, better than those archived by C4.5 in its original version in the same situations. In average, the model was reduced in 78.2%.

VII. CONCLUSIONS

When petroleum data are analyzed by means of data mining, machine learning or soft computing techniques oriented to classification, it is often the case that we must confront a discretization or a feature selection process, or both.

In this paper, we proposed a method to discretize and to select a subset of relevant features. The main contribution is to use the same criterion applied in the induction process to discretize and to select the relevant features. The results obtained (as measured by means of C4.5 error) suggest that this may be a better way of tackling these important problems. We suggest that applying discretization and afterwards feature selection, give us better results that the opposed sequence. It has some advantages: there is a loss of information when the discretization is made (e.g. different values are consider as the same class). When a posterior feature selection process is realized, this loss of information is taken into account.

When continuous feature selection algorithms are used, sometimes it is difficult to choose the number of features to be selected. In this way, previous discretization can help us to have a bias about this selection. Briefly, when some discretized attribute contains only one value, it can be considered as irrelevant [13]. In the same way, we observed that Relief (*weighing*) and SFFS–C4.5 (*subset*) can give us interesting advantages. When both algorithms are used, SFFS–C4.5 helps us to choose the number of attributes that better suits Relief and Relief lets us to know the contribution of each feature over the data set.

For comparison purposes, in the experiments carried out through this work, we have applied the same parameters and configuration to all data sets and algorithms; however, when carrying out further experiments we found out that if we personalize the parameters the results given by ES–C4.5 can be pretty improved.

Finally, we would like to mention that during these experiments we observed that the final errors given by ES–C4.5 and SFFS–C4.5 were, most of the times, better than those achieved by C4.5 alone in its in the same situations.

VIII. ACKNOWLEDGMENTS

The authors are very grateful to Hector Alfredo Carmona Pozos, Luis Carlos Fernández and Ernesto Carbajal Rivera for give us the geological, bio–technological and the computational support, respectively. This work was supported by the Instituto Mexicano del Petróleo and the Spanish project CICyT DPI2002-03225.

REFERENCES

[1] T. Back. *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, 1996.

[2] R. A. Caruana and D. Freitag. How Useful is Relevance? In *Proc. of the Fall'94 AAAI Symposium on Relevance*, New Orleans, 1994.

[3] J. Catlett. On Changing Continuous Attributes into Ordered Discrete Attributes. In *Proc. of the European Working Session on Learning*, pages 164–178, Berlin, Germany, 1991. Springer Verlag.

[4] J. Doak. An Evaluation of Feature Selection Methods and their Application to Computer Security. Technical Report CSE–92–18, Univ. of California, 1992.

[5] U. M. Fayyad and K. B. Irani. Multi–interval Discretization of Continuous–valued Attributes for Classification Learning. In *Proc. of the 13th Int. Conf. on Machine Learning*, pages 1022–1027. Morgan Kaufmann, 1993.

[6] R. Kerber. ChiMerge: Discretization of Numeric Attributes. In *Proc. of the 10th Nat. Conf. on Artificial Intelligence*, pages 123–127. The MIT Press, 1992.

[7] K. Kira and L. Rendell. A Practical Approach to Feature Selection. In *Proc. of the 9th Int. Conf. on Machine Learning*, pages 249–256, Aberdeen, Scotland, 1992. Morgan Kaufmann.

[8] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, London, U. K., 1998.

[9] M. Nikravesh and F. Aminzadeh. Past, Present and Future Intelligent Reservoir Characterization. *Journal of Petroleum Science and Engineering*, 31:67–79, 2001.

[10] PEMEX. *El Petróleo*. PEMEX, México, 1984.

[11] P. Pudil, J. Novovicová, and J. Kittler. Floating Search Methods in Feature Selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994.

[12] J. R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[13] J. J. Valdés, L. C. Molina, and N. Peris. Simultaneous Supervised Discretization of Numeric Attributes: a Soft Computing Approach. In *Proc. of the Congress on Evolutionary Computation*, pages 151–156, Camberra, Australia, 2003.

[14] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2001.