





Measuring Features Strength in Probabilistic Classification

Rosario Delgado¹  and Xavier-Andoni Tibau² 

¹ Department of Mathematics, Universitat Autònoma de Barcelona, Edifici C- Campus de la UAB. Av. de l'Eix Central s/n., 08193 Bellaterra (Cerdanyola del Vallès), Barcelona, Spain
delgado@mat.uab.cat

² Institute for Data Science, German Aerospace Center, 07745 Jena, Germany
xavier.tibau@dlr.de

Abstract. Probabilistic classifiers output a probability of an input being a member of each of the possible classes, given some of its feature values, selecting most probable class as predicted class. We introduce and compare different measures of the feature strength in probabilistic confidence-weighted classification models. For that, we follow two approaches: one based on conditional probability tables of the classification variable with respect to each feature, using different statistical distances and a correction parameter, and the second one based on accuracy in predicting classification from evidences on each isolated feature. On a case study, we compute these feature strength measures and rank features attending to them, comparing results.

Keywords: Probabilistic classifier · Feature strength
Statistical distance · Prediction accuracy

1 Introduction

In machine learning and statistics, classification is the problem of identifying to which of a set of classes or categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations whose classification membership is known. Any individual observation is analyzed into a set of quantifiable properties or characteristics, termed features, from which its category or class is to be predicted. In this work we allow features to be binary, categorical or discrete. An algorithm that implements classification, mapping input data to a category or class (output) is known as a classifier. A good classifier is one that predicts that output accurately.

A common subclass of classification is *probabilistic classification*. Algorithms of this nature use statistical inference to find the best class for a given instance. Unlike other algorithms, which simply output a “best” class, probabilistic algorithms output a probability of the instance being a member of each of the possible classes. The best class is then selected as the one with the highest probability,

and is the “predicted class” for the input. Such algorithms have advantages over non-probabilistic classifiers. Among them, it can output a confidence value associated with its choice and, correspondingly, it can abstain when its confidence of choosing any particular output is too low; in this way, it allows to adapt both sensitivity and specificity of the model depending on priorities.

Bayesian classifiers are probabilistic classification procedures that provide a natural way of taking into account any available information about the composition of the sub-populations associated with the different groups within the overall population. If the elements of the population are grouped into sub-populations or classes because they have common values of their features, then it could be natural to try to predict the values of the features or attributes for the members of any fixed class. On the opposite, if the class is unknown, Bayes’ rule can be used to predict the class given (some of) the feature values. The Bayesian classifier is a probabilistic model including the class variable and the features, and perhaps other (latent) variables as well. This model, after construction and validation, can be used to predict (infer) the classification of any new element. The simplest case is the naive Bayesian classifier, which makes the assumption that the features are conditionally independent of each other given the classification, but other models can be considered. We will assume that the considered features, as well as the class variable, are all discrete. Although many classification methods have been developed specially for binary classification, we do not restrict ourselves to this scenario but extend our study to the multiclass setting. More specifically, we consider a probabilistic classifier for class variable C with feature variables F_1, \dots, F_r .

To build effective models, data as accurate as possible is needed, but in real life, with limited resources, obtaining accurate data could have a huge associated cost. In this context, the need to decide on what features we focus on is clear. At the very first steps of the process, this is solved by passing the sieve of feature selection techniques (see Friedman et al. [2]), which attempt to shrink the dimensionality of the dataset to improve both accuracy (e.g. by avoiding overfitting or reducing variance), and interpretability. But later on, once the model is built, analyzing the importance of each feature is still a matter of importance not only because of the comments above, but because in a certain way, a model is a simplistic approximation of reality, so knowing what features are revealed as fundamental in a given model may be a clue about what features we should zoom in when looking for a deeper knowledge of the phenomenon. That is, feature strength focuses on the interpretability of the model and not on its simplification or reduction. After feature selection and the construction of the model of the desired dimension, it is a deeper step in the interpretation of the model, in which it is intended to analyze the influence of the features in the classifier.

Our goal is to introduce and compare different measures of the features strength for classification the classifier. As far as we know, there are no precedents on this type of study. There are, however, some works related in a certain sense. Indeed, different software packages deal with the question of measure the

strength of influence between neighboring nodes in a Bayesian network through some kind of strength measure of the arcs among them in the Directed Acyclic Graph (DAG). Only to mention two of them:

1. SMILE (Structural Modeling, Inference, and Learning Engine) is a fully platform independent portable library of C++ classes implementing graphical decision-theoretic methods, such as Bayesian networks, and influence diagrams and structural equation models. Among its tools, we find the *strength of influence* tool of a directed arc, which is always calculated from the CPT of the child node and essentially expresses some form of distance between the probability distributions of the child node conditional on the state of the parent node. With respect to this work, we introduce two families of measures of influence features in classification, not only these which are children nodes of the class. The first one is of the type considered in GeNIe, while the second one is inherently different in nature.
2. Similar to the *strength of influence* of SMILE is the *magnitude of influence* that the software Elvira computes. Paper [5] introduces the *magnitude of influence* of a link (MI) of Elvira for ordinal variables. This definition has no sense if variables are discrete but not ordered. We introduce measures of influence named *strength measures*, which apply for discrete variables (ordinal or not), and not only refer to the directed arc from a parent to a child, but they apply to any pair of variables.

We have followed two different approaches to the problem of defining measures of the strength of a feature in a probabilistic classifier: one based on conditional probability tables (CPT) of the classification variable C to the feature, using different statistical distances, and the other based on accuracy in prediction. With respect to the first one, which is the subject of Sect. 2, statistical distances, divergences, and similar quantities have a large history and play a fundamental role in statistics, machine learning and associated scientific disciplines. Statistical distances are defined in a variety of ways, by comparing probability mass distributions in the discrete probability models context, as is the case at hand. We will choose four of them as an example in our case study (Sect. 4). We will assign a measure of strength to each feature, say F_i , as the “*maximum discrepancy*” observed on the CPT of C conditioned to feature F_i , which is defined as the maximum distance in the pairwise comparisons corresponding to the conditional probability distribution of C to different fixed values of F_i .

In the second case, considered in Sect. 3, the strength measure of each feature F_i for the classification variable is defined as the corresponding accuracy when predicting classification from an evidence expressed exclusively in terms of F_i . In both cases we obtain a ranking of the features in classification (in the first case, a possible different ranking is obtained for each statistical distance that has been considered). In Sect. 4 we apply the different measures we have introduced in previous sections to a case study, and compare features rankings by using both Hamming distance and the degree of consistency.

2 Strength Measures Based on CPT

In this section we introduce some strength measures to deeply analyze to what extent class variable C is affected by the different features in the classifier. Fix a feature variable F . We would like to compute a measure of the effect of induced changes in F on the conditional probability distribution of variable C . For that, we consider evidences of the form $F = a$ and estimate from sample probabilities for the query variable C given the evidence, $P_a^F(x) = P(C = x / F = a)$ with $x \in \mathcal{C}$ (\mathcal{C} being the set of possible outcomes of variable C), if $P(F = a) > 0$. These probabilities conform the Conditional Probability Table (CPT) of C conditioned to F . We propose an approach based on a statistical distance.

Different statistical distances or divergence measures have been introduced in the literature between two discrete probability distributions. For example, and only to mention four of them, the Kullback-Leibler divergence, the Pearson chi-square distance, the Hellinger measure or the Kolmogorov distance (see [6]). Some of them are asymmetric, provoking that changing order of the arguments can yield substantially different values. For that, we consider symmetrized versions of them. We denote by KL the Kullback-Leibler distance or divergence, also known as *relative entropy*, that is, given two discrete probability distributions taking values $x \in \mathcal{X}$, Q_1 and Q_2 , and with the understanding that there is not $x \in \mathcal{X}$ such that $Q_1(x) = Q_2(x) = 0$, $KL(Q_1, Q_2) = \sum_{x \in \mathcal{X}} Q_1(x) (\log Q_1(x) - \log Q_2(x))$, with the convention that $0 \log(0) = 0$. Note that $KL(Q_1, Q_2) \geq 0$ although it could be $+\infty$ (if $Q_2(x) = 0$ for some x). Following [4], we symmetrize this distance by means of the harmonic sum, that is, the half the harmonic mean, of the component Kullback-Leibler divergences. Pearson chi-squared and Hellinger distances have also been symmetrized.

Distance name	Formula
Kullback-Leibler	$d_1(Q_1, Q_2) = 1 / \left(\frac{1}{KL(Q_1, Q_2)} + \frac{1}{KL(Q_2, Q_1)} \right)$
Pearson chi-squared	$d_2(Q_1, Q_2) = \sum_{x \in \mathcal{X}} 2 \frac{(Q_1(x) - Q_2(x))^2}{Q_1(x) + Q_2(x)}$
Squared blended Hellinger	$d_3(Q_1, Q_2) = \sqrt{\sum_{x \in \mathcal{X}} 2 (\sqrt{Q_1(x)} - \sqrt{Q_2(x)})^2}$
Kolmogorov-Smirnov	$d_4(Q_1, Q_2) = \max_{x \in \mathcal{X}} Q_1(x) - Q_2(x) $

Note that $d_i \geq 0$, but that if there exist $x_1, x_2 \in \mathcal{X}$ such that $Q_1(x_1) = 0$ and $Q_2(x_2) = 0$, then $d_1(Q_1, Q_2) = +\infty$.

We introduce a strength measure for feature F based on a statistical distance or symmetrised divergence measure d , which could be any of the previous distances d_1, \dots, d_4 , or even another, and name it *Strength Distance (SD)*, in this way:

$$SD(F) = \max_{a, b \in \mathcal{F}} d_{a, b}^F$$

where \mathcal{F} is the set of the possible outcomes of variable F , and $d_{a,b}^F$ denotes the statistical distance between P_a^F and P_b^F , that is, $d_{a,b}^F = d(P_a^F, P_b^F)$. Therefore, we compute strength distance from pairwise comparatives through the statistical distance or symmetrised divergence d .

Proposition 1. $SD(F) \geq 0$, and $SD(F) = 0$ if and only if F and C are independent.

Proof. $SD(F) \geq 0$ by definition. On the other hand, $SD(F) = 0$ if and only if $d_{a,b}^F = 0$ for any $a, b \in \mathcal{F}$, but this fact is equivalent to say that $P_a^F = P_b^F$ for any $a, b \in \mathcal{F}$, which is equivalent to the independency between F and C . \square

Note that although we have defined SD through the maximum, we could have chosen any other aggregation function of the distances $d_{a,b}^F$ that verified Proposition 1. The maximum is the least robust (jointly with the minimum) option, since it is maximally sensitive to extreme values, which represents an advantage if extreme values are real (not measurement errors), as in our case, where they are of great importance to assess the strength of a feature for classification.

Because previous measure does not consider if different instantiations of a feature variable produce different predictions for class variable C , it seems appropriate to introduce a correction that does take account of this fact.

Let $\alpha = \#C$ and $\beta(F) = \#\mathcal{F}$, where $\#$ denotes the cardinal of a finite set, and let $\gamma(F)$ denote the number of different predictions obtained from the classifier for C given the evidences $E = \{F = a\}$, with a varying in \mathcal{F} , that is,

$$\gamma(F) = \#\{\arg \max_{x \in C} P(C = x / F = a), a \in \mathcal{F}\}.$$

Then, define

$$\delta(F) = \frac{\gamma(F)}{\min(\alpha, \beta(F))} \in (0, 1],$$

which is the proportion of different predictions actually obtained by the classifier for class C among the possible we could obtain from an evidence on F . Therefore, $\delta(F)$ is a measure of the influence of feature F on C , and we can use it to correct strength measure SD by introducing the *Corrected Strength Distance (CSD)* in this way: $CSD(F) = SD(F) \times \delta(F)$, which is $\leq SD(F)$. Note that as SD , $CSD(F) \geq 0$, and $CSD(F) = 0$ if and only if F and C are independent.

3 Strength Measures Based on Accuracy in Prediction

In general, after constructing the classifier from the dataset, we perform validation of the model, and once validated, we can use it for future predictions. Validation consists of a procedure for assessing how the classifier performs in the sense of correctly predict the query variable C from any evidence given in terms of the features. The most elementary validation procedure is based on splitting the dataset into two parts, *training* and *test sets*, what is known as *split-validation*. Cross-validation procedure is one of the most widely used methods for

estimating prediction error, most common forms for implementation are *k-fold cross-validation* and its particular case ($k = r$) *leave-one-out cross-validation*. We will apply *leave-one-out cross-validation*, from which we obtain *accuracy*, defined as the success rate in prediction, that is, $\frac{\#\{Matches\}}{\#\{Validation\ set\}}$. In addition, for each fixed feature F , we can apply the procedure by predicting the class outcome from single evidences on F , and estimate the accuracy in predicting class C . We denote it by $Acc(F)$.

Nevertheless, this strength measure for each feature does not take into account the following fact: if feature F were *independent* of class variable C , $\gamma(F) = 1$ since for any $a \in \mathcal{F}$, $\arg \max_{x \in C} P(C = x / F = a) = Mode(C)$, with $Mode(C)$ the most frequent class in dataset, that is, if F and C were independent, prediction for C will always be its more likely outcome, independently of the instantiation of F . Denote by p_{mode} the relative frequency of this value in the dataset, $p_{mode} \geq 1/\alpha$. In general, $Acc(F) \geq 1/\alpha$ but if F and C were independent, we would have $Acc(F) = p_{mode}$. Therefore, it seems natural to scale *accuracy* and introduce the *Relative Increment in Accuracy (RIA)* (with respect to p_{mode}) of any feature F by

$$RIA(F) = \frac{Acc(F) - p_{mode}}{p_{mode}}.$$

Proposition 2. *RIA verifies the following properties:*

- (a) $RIA(F) = 0$ if F and C are independent, but the reciprocal is not true.
- (b) $-1 < c_1 - c_2 \leq RIA(F) \leq c_1 \leq \alpha - 1$, with $c_1 = \frac{1-p_{mode}}{p_{mode}}$, $c_2 = (c_1 + 1) \frac{\alpha-1}{\alpha}$.

Proof

- (a) If F and C are independent, the predicted class given any evidence on F will be always the same, $Mode(C)$. Therefore, the proportion of correct prediction, which is $Acc(F)$, has to be equal to p_{mode} by definition.
- (b) First, since $Acc(F) \leq 1$, $RIA(F) \leq \frac{1-p_{mode}}{p_{mode}} = c_1$, and $c_1 \leq \alpha - 1$ due to the fact that $p_{mode} \geq 1/\alpha$. Secondly, since $Acc(F) \geq 1/\alpha$, we have that

$$RIA(F) \geq \frac{1/\alpha - p_{mode}}{p_{mode}} = \frac{(1/\alpha - 1) + (1 - p_{mode})}{p_{mode}} = c_1 - \frac{1 - 1/\alpha}{p_{mode}}$$

and $c_2 = \frac{1-1/\alpha}{p_{mode}}$ can be written as $c_2 = (c_1 + 1) \frac{\alpha-1}{\alpha}$ if we use that by definition of c_1 we can isolate and obtain $p_{mode} = \frac{1}{c_1+1}$. □

Interpretation of $RIA(F) < 0$ is that F is a feature that as predictor is worse than choosing the most common class. That is, to make classification, it is worst to use evidence on F than nothing, just the opposite that if $RIA(F) > 0$, case in which the higher the value of $RIA(F)$, the stronger the influence of feature F in classification. Therefore, this measure allows to make a ranking of the features, taking into account the strength of their influence in the classification process.

Particular Cases:

- (i) Uniform distribution of class C in the dataset. Then, $p_{mode} = 1/\alpha$ and $c_1 = c_2 = \alpha - 1$, obtaining $0 \leq RIA(F) \leq \alpha - 1$.
- (ii) Binary classification ($\alpha = 2$). Then, $c_2 = (c_1 + 1)/2$ and $-1 < \frac{c_1-1}{2} \leq RIA(F) \leq c_1 \leq 1$.
- (iii) If our situation is a combination of both, that is, binary classification and uniform distribution of C into the database, then $0 \leq RIA(F) \leq 1$.

4 Case Study

We consider a dataset of 1,597 policing clarified arson-caused wildfires (for which the alleged offenders have been identified), that has been feeding since 2008 by

Table 1. Variables in the dataset of the arson-caused wildfires.

Forest fire features	Outcomes
$C_1 =$ season	Spring/winter/summer/autumn
$C_2 =$ risk level	High/medium/low
$C_3 =$ start time	Morning/afternoon/evening
$C_4 =$ starting point	Pathway/road/houses/crops/interior/forest track/others
$C_5 =$ use burned surface	Agricultural/forestry/ livestock/interface/recreational
$C_6 =$ number of seats	One/more
$C_7 =$ related offense	Yes/no
$C_8 =$ pattern	Yes/no
$C_9 =$ traces	Yes/no
$C_{10} =$ who denounces	Guard/particular/vigilance
Arsonist characteristics	Outcomes
$A_1 =$ age	$\leq 34/35-45/46-60/>60$
$A_2 =$ way of living	Parents/in couple/single/others
$A_3 =$ kind of job	Handwork/qualified
$A_4 =$ employment status	Employee/unemployed/sporadic/retired
$A_5 =$ educational level	Illiterate/elementary/middle/upper
$A_6 =$ income level	High/medium/low/without incomes
$A_7 =$ sociability	Yes/no
$A_8 =$ prior criminal record	Yes/no
$A_9 =$ history subst. abuse	Yes/no
$A_{10} =$ history psychol. probl.	Yes/no
$A_{11} =$ stays in the scene	No/remains there/remains and gives aid
$A_{12} =$ distance home-scene	Short/medium/long/very long
$A_{13} =$ displacement means	On foot/by car/all terrain/others
$A_{14} =$ residence type	Village/house/city/town
$A_{15} =$ motivation (Class)	Slight negligence/gross negligence/impulsive/profit/revenge

the Secretary of State for Security throughout the entire Spanish territory, under the leadership of the Prosecution Office of Environment and Urbanism of the Spanish state, and contains information obtained from a specific questionnaire concerning authors that have been arrested or imputed. This dataset is an update of that considered in [1]. A total number of $n = 25$ categorical variables are consigned, from which 10 refer to forest fire features, C_1, \dots, C_{10} , and the rest to arsonist characteristics, A_1, \dots, A_{15} , and are described in Table 1.

Bayesian network classifier has been constructed from the dataset with the restriction that directed arcs from forest fire features to arsonist characteristics are forbidden, and is used for classification variable A_{15} , which is author motivation and has proved to be the most significant author variable (see [1]) and forest fire features C_1, \dots, C_{10} .

All calculations, as well as the process of model construction, validation and inference, have been carried out with **R** (<https://cran.r-project.org>). Two packages of R has been adopted: bnlearn, for network and parameter learning, and gRain, for making inference by probability propagation.

4.1 Ranking Features by Strength Using Measures Based on CPT

The CPT of class variable A_{15} with respect to any of the features C_1, \dots, C_{10} , are learned from the dataset and given in Tables 6, 7, 8, 9, 10 and 11 in the Appendix. Fixed the evidence in terms of a feature variable and one of its values (that is, fixing a column in a CPT), the corresponding predicted class is the most likely, that is, that with the highest probability, which is highlighted in boldface.

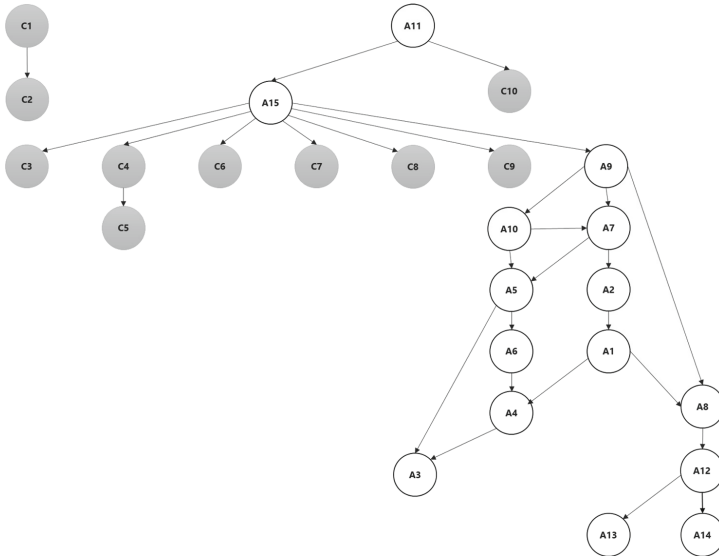


Fig. 1. Learned structure (DAG) of the BN from the dataset of arson-caused wildfires.

Table 2. SD and CSD for the feature variables, using different distances.

Feature	$SD(C_i)$				$\delta(C_i)$	$CSD(C_i)$			
	d_1	d_2	d_3	d_4		d_1	d_2	d_3	d_4
C_3	0.2147	0.7506	0.6430	0.3892	2/3	0.1431	0.5004	0.4287	0.2595
C_4	0.4748	1.3499	0.9584	0.3603	2/5	0.1899	0.5399	0.3833	0.1441
C_5	0.0205	0.0801	0.2028	0.0821	1/5	0.0041	0.0160	0.0406	0.0164
C_6	0.2235	0.8069	0.6611	0.3181	2/2	0.2235	0.8069	0.6611	0.3181
C_7	0.5368	1.6341	0.9956	0.4470	2/2	0.5368	1.6341	0.9956	0.4470
C_8	0.2508	0.8646	0.7022	0.2597	1/2	0.1254	0.4323	0.3511	0.1299
C_9	0.0561	0.2073	0.3348	0.1238	1/2	0.0281	0.1037	0.1674	0.0619
C_{10}	0.0520	0.2041	0.3217	0.2233	2/3	0.0347	0.1361	0.2145	0.1488

Table 3. Ranking of the feature variables by SD and by CSD, using different distances, from the strongest (top) to the weakest (bottom).

Ranking by SD				Ranking by CSD			
d_1	d_2	d_3	d_4	d_1	d_2	d_3	d_4
C_7	C_7	C_7	C_7	C_7	C_7	C_7	C_7
C_4	C_4	C_4	C_3	C_6	C_6	C_6	C_6
C_8	C_8	C_8	C_4	C_4	C_4	C_3	C_3
C_6	C_6	C_6	C_6	C_3	C_3	C_4	C_{10}
C_3	C_3	C_3	C_8	C_8	C_8	C_8	C_4
C_9	C_9	C_9	C_{10}	C_{10}	C_{10}	C_{10}	C_8
C_{10}	C_{10}	C_{10}	C_9	C_9	C_9	C_9	C_9
C_5	C_5	C_5	C_5	C_5	C_5	C_5	C_5

In this way we see in Table 6, for example, that regardless of the value of feature C_1 , the prediction for A_{15} is always “slight negligence”, which is consistent with the fact that both are independent variables, as is deduced from the fact that in the DAG they appear as disconnected (see Fig. 1). Instead, evidences on feature C_3 can lead to predict “slight negligence” or “gross negligence”, depending on if C_3 = “morning” or “afternoon”, or C_3 = “evening” (see Table 7). In Table 2 we have the values of SD and CSD for the features $C_3 - C_{10}$ and the four statistical distances introduced in Sect. 2. Both are zero for C_1 and C_2 , since they are disconnected from A_{15} .

Glancing at Table 3, we realize that rankings of the features do not match for all the distances, although C_7 is the number one, and C_5 is always at the end of the classification. But if we restrict ourselves to CSD, C_6 is the top second for the four considered distances, while C_9 is the bottom second one.

4.2 Ranking Features by Strength Using Measures Based on Accuracy in Prediction

We perform *leave-one-out cross-validation* and the accuracy value $Acc(C_i)$ is obtained by dividing the number of correct predictions using as evidence the value of C_i , by the total number of predictions (excluding blanks). Both Acc values and RIA are recorded in Table 4. Take into account that p_{mode} is the probability of the most likely non-missing class in the dataset, normalizing probability after eliminating missing values. In this case, $p_{mode} = 697/1463 \simeq 0.47642$.

Finally, we compare rankings obtained from SD and CSD and that obtained by applying RIA criterion, by using both the Hamming distance and the degree of consistency indicator c (see [3]), as consigned in Table 5. In information theory, the Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different, that is, it measures the minimum number of substitutions required to change one string into the other. For two measures f and g on a domain Ψ , let $R = \{(a, b) \in \Psi \times \Psi : f(a) > f(b), g(a) > g(b)\}$ and $V = \{(a, b) \in \Psi \times \Psi : f(a) > f(b), g(a) < g(b)\}$. Then, the *degree of consistency* c of f and g is $c(f, g) = \frac{|R|}{|R|+|V|}$, where $|A|$ denotes the number of elements of the (finite) set A . We apply this indicator with f and g ranking functions. We observe that Hamming distance is minimized with CSD and distances $d_3 =$ Squared blended symmetric Hellinger distance, and

Table 4. Acc and RIA for the feature variables, which have been ranked from top to bottom in descending order.

Feature	$Acc(C_i)$	$RIA(C_i)$
C_7	0.4990	0.0474
C_{10}	0.4949	0.0387
C_3	0.4919	0.0316
C_4	0.4846	0.0172
C_6	0.4826	0.0129
C_8	0.4764	0.0000
C_9	0.4764	0.0000
C_5	0.4764	0.0000

Table 5. Hamming distance and degree of consistency indicator c between SD and CDS, with distance d_i , and RIA.

	Hamming SD-RIA	Hamming CSD-RIA	$c(\text{SD}, \text{RIA})$	$c(\text{CSD}, \text{RIA})$
d_1	5	5	19/28	21/28
d_2	5	5	19/28	21/28
d_3	5	3	19/28	22/28
d_4	5	3	24/28	24/28

$d_4 =$ Kolmogorv-Smirnov, while d_4 maximizes the consistency indicator, being CSD more consistent with RIA than SD. This reinforces the hypothesis that the correction in SD obtained multiplying by factor δ , improves it.

5 Conclusion

We introduce different measures of features strength in a probability classifier. From them, Corrected Strength Distance (CSD), which is based on CPT of class conditioned to each feature, seems to outperform Strength Distance (SD) since it is more consistent with Relative Increment in Accuracy (RIA), which is a measure based on accuracy in prediction. From the chosen distances, the best options have been Hellinger and Kolmogorov-Smirnov, both after correction.

Acknowledgments. The authors are supported by Ministerio de Economía y Competitividad, Gobierno de España, project ref. MTM2015 67802-P, and belong to the “Quantitative Methods in Criminology” research group of the Universitat Autònoma de Barcelona. They wish to express their acknowledgment to the Secretary of State for Security and the Prosecution Office of Environment and Urbanism of the Spanish state, for providing dataset used in the case study.

Appendix: Conditional Probability Tables of Features with Respect to Class Variable A_{15}

Table 6. CPT of A_{15} conditioned to C_1 (in %).

$C_1 \rightarrow$	Spring	Summer	Autumn	Winter
Pulsional	10.05	10.05	10.05	10.05
Gross negligence	31.31	31.31	31.31	31.31
Slight negligence	47.64	47.64	47.64	47.64
Profit	7.59	7.59	7.59	7.59
Revenge	3.42	3.42	3.42	3.42

Table 7. CPT of A_{15} conditioned to C_2 , and conditioned to C_3 (in %).

	$C_2 \downarrow$			$C_3 \downarrow$		
	High	Medium	Low	Morning	Afternoon	Evening
Pulsional	10.05	10.05	10.05	11.04	7.33	25.82
Gross negligence	31.31	31.31	31.31	19.63	33.18	30.22
Slight negligence	47.64	47.64	47.64	57.06	51.07	18.13
Profit	7.59	7.59	7.59	10.43	6.44	12.09
Revenge	3.42	3.42	3.42	1.84	1.97	13.74

Table 8. CPT of A_{15} conditioned to C_4 (in %).

$C_4 \rightarrow$	Pathway	Road	Houses	Crops	Interior	F. Track	Others
Pulsional	20.00	36.94	3.90	0.90	6.22	16.38	6.12
Gross negligence	24.19	16.22	32.47	35.75	33.97	26.72	35.61
Slight negligence	33.49	26.13	59.74	59.50	50.24	30.17	51.80
Profit	14.88	10.81	1.30	3.62	8.61	15.52	5.04
Revenge	7.44	9.91	2.60	0.23	0.96	11.21	1.44

Table 9. CPT of A_{15} conditioned to C_5 (in %).

$C_5 \rightarrow$	Agricultural	Forestry	Livestock	Interface	Recreational
Pulsional	6.28	12.16	11.88	13.40	10.58
Gross negligence	33.00	30.26	30.60	29.44	31.55
Slight negligence	52.51	44.30	44.61	45.96	45.88
Profit	6.13	8.96	8.81	6.96	7.92
Revenge	2.09	4.32	4.10	4.24	4.08

Table 10. CPT of A_{15} conditioned to C_6 , to C_7 and to C_8 (in %).

	$C_6 \downarrow$		$C_7 \downarrow$		$C_8 \downarrow$	
	One	More	Yes	No	Yes	No
Pulsional	7.77	25.13	51.39	7.72	27.57	3.65
Gross negligence	32.31	21.99	19.44	31.92	18.48	34.95
Slight negligence	52.23	20.42	5.56	50.26	28.74	54.71
Profit	5.18	23.04	6.94	7.65	19.65	3.65
Revenge	2.51	9.42	16.67	2.45	5.57	3.04

Table 11. CPT of A_{15} conditioned to C_9 and to C_{10} (in %).

	$C_9 \downarrow$		$C_{10} \downarrow$		
	Yes	No	Guard	Vigilance	Particular
Pulsional	6.55	12.11	11.69	12.08	8.85
Gross negligence	34.06	29.58	37.39	41.17	26.18
Slight negligence	55.68	43.30	38.26	33.09	55.41
Profit	2.40	10.61	8.69	9.22	6.65
Revenge	1.31	4.39	3.96	4.44	2.90

References

1. Delgado, R., González, J.L., Sotoca, A., Tibau, X.-A.: A Bayesian network profiler for wildfire arsonists. In: Pardalos, P.M., Conca, P., Giuffrida, G., Nicosia, G. (eds.) MOD 2016. LNCS, vol. 10122, pp. 379–390. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51469-7_31
2. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. Springer Series in Statistics, vol. 1, pp. 337–387. Springer, New York (2001). <https://doi.org/10.1007/978-0-387-21606-5>
3. Huang, J., Ling, C.: Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng. **17**, 299–310 (2005). <https://doi.org/10.1109/TKDE.2005.50>
4. Johnson, D., Sinanovic, S.: Symmetrizing the Kullback-Leibler distance. Technical report, ECE Publications, Rice University (2001). <https://www.ece.rice.edu/~dhj/resistor.pdf>
5. Lacave, C., Luque, M., Díez, F.J.: Explanation of Bayesian networks and influence diagrams in Elvira. IEEE Trans. Syst. Man Cybern.-Part B: Cybern. **37**(4), 952–965 (2007). <https://doi.org/10.1109/TSMCB.2007.896018>
6. Markatou, M., Chen, Y., Afendras, G., Lindsay, B.G.: Statistical distances and their role in robustness. In: Chen, D.-G., Jin, Z., Li, G., Li, Y., Liu, A., Zhao, Y. (eds.) New Advances in Statistics and Data Science. IBSS, pp. 3–26. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69416-0_1