**Proceedings of the 2007 IEEE Symposium on
Computational Intelligence and Data Mining (CIDM 2007)**

Validity of Probabilistic Rules

Marina Sapir, Mikhail Teverovskiy
Aureon Labs,
28 Wells Ave, Yonkers, USA

***Abstract*— We propose an axiomatic approach to defining of the validity of probabilistic inductive rules $E \Rightarrow H$. The set of rules is evaluated against an available dataset, where the conditions E, H are either true or false for each instance in the dataset. Introduced here are six axioms which formalize common sense dependencies between the validity of rules and their support, confidence, lift and amount of available evidence. Having a single validity measure, contrary to multiple criteria, helps compare and rank induced rules. We demonstrate that the z-test of difference of proportions satisfies all the axioms and can be used as a measure of rules validity. Knowing that the z-test statistics is normally distributed, allows one to filter out statistically unreliable rules. We demonstrate advantages of the proposed approach on a real life medical dataset.**

I. Introduction

Due to insufficient information for induction, the rules inducted from a set of data are not always supposed to be true. Therefore, the question arises, how to evaluate the "validity" of uncertain rules; how to find "interesting", important rules among all possible ones.

Consider, for example, one of the most popular approaches: association rules [1]. "Interesting" association rules are defined by thresholds on support and confidence. The thresholds have to be selected before the analysis. However, the level of association on the given dataset is a fundamental property of the data, which can be discovered only during the analysis.

As an instructive example, we consider a paper [6], which proposes a new way to trim some redundant association rules. To demonstrate the advantages of the method, the author applies it on several datasets. The selected for these datasets support levels are ranging from 0.5%, to 97%. Yet, there is no hint in the paper, how these thresholds may be selected on new datasets before the analysis. If the threshold selection is not data-specific, the analysis will produce either too few or too many rules, some of which are not reliable at all.

Similar problems arise in other data mining approaches. For example, the probabilistic rough set approach proposed in [2] requires setting up two *certainty control* limits for the inclusion of elementary sets in the set approximation, as well as a *certainty gain* threshold for the positive and boundary regions.

To avoid setting the data-specific thresholds a priori, one may rank inducted rules by their validity and select only the "top" rules [7]. Here we propose a general approach to the problem of rules evaluation. We introduce six axioms, which characterize the relationships between the "validity" and the empirical criteria "confidence", "support" and "lift", traditionally used to evaluate rules.

We show that the known z-test of difference of proportions can be used as a validity function satisfying all the axioms. Instead of setting a threshold for selecting "interesting" rules, the z-test allows one to discard statistically insignificant rules. The advantage of this procedure is that one can select the same level of significance for all analyzed datasets.

II. Axiomatic definition of validity

The goal is to formulate general commonly acceptable epistemological requirements for validity of production rules.

As a starting point, we can follow (with some modifications) the four axioms already formulated for the "quality functions" in [4] with reference to earlier works.

The axioms for the validity of rules are formulated here for an arbitrary rule $(E \Rightarrow H)$ where E, H are conditions on the characteristics of instances. Alternatively, we will be talking about events E, H, implying that event E (or H) occurs when the conditions E (or H) on an instance' characteristics are satisfied.

We rephrase the axioms, postulating relationships between the validity of a rule and popular empiric criteria *support, confidence* and *lift*. For convenience, the criteria are defined below.

For a rule $E \Rightarrow H$,

$$\textit{confidence}(E \Rightarrow H) = P(H|E),$$

the conditional probability of the event H under conditions of event E;

$$\textit{support}(E \Rightarrow H) = P(E),$$

the probability of the premise E;

$$\textit{lift}(E \Rightarrow H) = \frac{P(H|E)}{P(H)},$$

the ratio of conditional and unconditional probabilities of the event H.

If lift is less than one, the event E decreases the chances of the event H. All the axioms describe validity only for rules with a lift larger than one. If lift is less than one, the validity may be expressed by negative a numbers, however we will not make such a requirement here.

Now, the axioms about the validity function $Q(E, H)$ on rules $E \Rightarrow H$ can be expressed as follows.

1) $Q(E, H) = 0$**, when lift equals 1.**
   One would expect that an ability to predict a conclusion by a premise is lowest when the premise and the conclusion are not related at all. This is the case when lift

equals 1, and the premise E does not affect the chances of conclusion H. The non-negative validity function shall be equal to 0 in this case.

2) $Q(E, H)$ **increases in confidence of rules for a fixed support**
   If two rules have equal support, the rule with higher confidence is more valid.
3) $Q(E, H)$ **increases in support of rules for a fixed confidence.**
   If two rules have equal confidence, the rule with higher support is more valid.
4) $Q(E, H)$ **decreases in support of rules with constant product of support and confidence.**
   This axiom formalizes an intuition that confidence is more important than support for the validity of rules. If two rules have equal product of support and confidence, the rule with higher confidence but lower support is more valid.

   We would like to postulate two additional properties describing the dependence of a validity criterion on size of a dataset, where rules are evaluated, and on the "lift" of a rule.

5) $Q(E, H)$ **increases in size n of a dataset for a fixed confidence, support of rules and unconditional probability $P(H)$.**
   All previous axioms compare only rules evaluated against the same data. Yet, as available evidence grows, our reliance on the rule grows too.
6) $Q(E, H)$ **increases in lift of rules for fixed confidence, support and size n of a dataset.**
   The axiom states that knowing support and confidence of a rule and dataset's size is not sufficient to determine the validity of the rule. One needs to know how much knowledge about the conclusion we can gain from the premise. Lift is a measure of association between the premise and conclusion on the given dataset, and it is important for evaluation of the rule's validity.

The axioms summarize general requirements for the validity criterion of inductive rules. We will call validity criteria, satisfying all the axioms, ***justified***.

III. An example of a justified validity criterion

We propose an example of a justified validity criterion. It is the z-test (without the correction for continuity) [5], evaluating two sided hypothesis that the unconditional probability of conclusion $p = P(H)$ and the conditional probability $p_1 = P(H|E)$ are equal.

Let n denotes the size of the dataset; n_1 the number of cases with the event E in the dataset, $q = 1 - p$. The criterion may be presented in such form:

$$z(p_1, n_1, p, n) = \frac{p_1 - p}{\sqrt{(p\,q)/n_1}}. \qquad (1)$$

In our notations, the confidence of the rule is p_1, the support of the rule is n_1/n, the lift is p_1/p.

Let us show that all the axioms hold for this criterion.

Theorem 3.1: The z-test of difference of proportions is a justified criterion for the validity of inductive rules.

Proof. We need to demonstrate that all the axioms are true for the z-test (1).

1) We need to prove that if $p_1/p = 1$, $z(p_1, n_1, p, n) = 0$. This is obvious: if $p_1 = p$, the function (1) equals 0.
2) We need to prove that on a given dataset, for fixed support (n_1/n), the proposed validity function increases with confidence p_1 of the rules. Indeed, when (n_1/n) and n, p are fixed, and p_1 increases, so does the value of the function z.
3) We need to prove that on a given dataset, for the fixed confidence p_1, the validity function (1) increases with the support (n_1/n). Indeed, for fixed p, n, when the support increases, the value n_1 increases. With fixed confidence p_1, the numerator of the formula (1) stays the same, and the denominator decreases. Therefore, the function z increases.
4) We need to show that, for a given dataset, the function (1) decreases in support (n_1/n) for a fixed product of support and confidence: $p_1 \cdot (n_1/n)$.
   To see it, let us multiply both the numerator and the denominator of the expression (1) for the function z by n_1. With fixed n, n_1 is proportional to support of the rule.
   In the numerator, we will have the expression
   $$p_1\ n_1 - p\ n_1.$$
   The first product is constant, when product of support and confidence is constant. The second product is increasing when the support (n_1/n) is increasing on the given dataset. Therefore, the numerator is decreasing when support is increasing.
   The denominator
   $$n_1\sqrt{(p\ q)/n_1} = \sqrt{p\ q\ n_1}$$
   increases with support (n_1/n). Therefore whole expression for the function z decreases with support (n_1/n) for the fixed product of support and confidence.
5) We need to show that the proposed validity function increases in n for fixed support (n_1/n), confidence p_1 and unconditional probability p.
   If confidence p_1 and support (n_1/n) are fixed, the number n_1 increases with the growth of n. Therefore, the denominator of the formula (1) decreases, while the numerator stays constant, and the function z increases.
6) Now, we need to demonstrate that the proposed validity function increases with lift $l = (p_1/p)$ for the fixed confidence p_1 and support (n_1/n).
   To do it, we need to transform the formula (1) to show how the function depends on the lift. We assume $l > 1$. We will replace p with $\frac{1}{l}p_1$, transforming both the

numerator and the denominator of the expression for the function z. For the numerator we have:

$$p_1 - p = p_1 - p_1/l = \frac{p_1}{l}(l-1). \quad (2)$$

For the denominator we have:

$$\sqrt{\frac{p\,q}{n_1}} = \sqrt{\frac{(p_1/l)\cdot(1-p_1/l)}{n_1}} = \sqrt{\frac{p_1 l - p_1^2}{l^2 n_1}} = \\ = \frac{1}{l}\sqrt{\frac{p_1(l-p_1)}{n_1}}. \quad (3)$$

Now, if we multiply the numerator (2) and the denominator (3) by l/p_1, the expression (1) will take the following form:

$$\frac{l-1}{\sqrt{(l-p_1)/(p_1 n_1)}} = \frac{l-p_1+p_1-1}{\sqrt{(l-p_1)/(p_1 n_1)}} = \\ = \frac{l-p_1}{\sqrt{(l-p_1)/(p_1 n_1)}} - \frac{1-p_1}{\sqrt{(l-p_1)/(p_1 n_1)}} = \\ = \frac{\sqrt{l-p_1}}{\sqrt{1/(p_1 n_1)}} - \frac{1-p_1}{\sqrt{(l-p_1)/(p_1 n_1)}} \quad (4)$$

When confidence p_1, support n_1/n and the size n of the dataset are fixed, the first nonnegative part of the expression (4)

$$\frac{\sqrt{l-p_1}}{\sqrt{1/(p_1 n_1)}}$$

increases in l, the second nonnegative part of the expression (4)

$$\frac{1-p_1}{\sqrt{(l-p_1)/(p_1 n_1)}}$$

decreases in l, therefore the whole expression (4) increases in l.

This proves that the z-test satisfies all requirements for the criterion of the validity of rules. **Q.E.D.**

IV. Rules filtering using z-test

The z-test has an additional advantage: for large samples, it has normal distribution when the true proportion of the event C is identical in the whole population and under conditions E of the rule $E \Rightarrow C$.

If we select an acceptable significance level α, it sets the lowest value of the z-test t_α, which allows one to reject the zero-hypothesis about equality of proportions p, p_1. This is a natural threshold to filter out statistically unreliable rules: all rules with the z-test value below the threshold value, can be considered invalid, unreliable.

Note that the z-test can not be used for selecting significant rules, because of the multiple testing problem. We test many rules on a finite dataset, and by chance some of them will have high values of the z-test on the data even if premise and conclusion are independent in general population. An advantage of rule-based approach is that an area expert can have a final word in filtering out spurious rules with high statistical significance.

V. Using a justified validity criterion on a real life data

We want to demonstrate on a real life dataset advantages of rules' selection with a justified validity criterion (the z-test) versus the traditional approach, when multiple measures with fixed thresholds are used.

The data represent information about prostate cancer patients from the Memorial Sloan-Kettering Cancer Center. The goal is to predict clinical failure of a patient (death due to the disease or metastases) during five years after prostatectomy. The patients are characterized by 17 features. The information includes clinical characteristics, some measurements of abundance of androgen receptor in the prostate tissue, as well as histological properties of the tissue, identified during computerized analysis of H & E images. The dataset was randomly split on the training (295 records) and test data (288 records). The clinical failure class (the first class) makes only 7.8% of the training set, and only 5.9% of the test set. Most of features are continuous.

On the training data, we applied two algorithms, generating interval production rules

$$(a_1 \le x_1 \le b_1)\&\ldots\&(a_m \le x_m \le b_m) \Rightarrow y = c.$$

The first algorithm [8] uses traditional threshold-based approach for the rules evaluation. It finds the most general interval rules, satisfying given constraints on support and confidence. The rule $E_1 \Rightarrow H$ is ***more general*** than the rule $E_2 \Rightarrow H$, if $E_2 \vdash E_1$.

The "concentration algorithm" [9] uses the z-test as a single validity criterion. The algorithm finds "digest of rules": compact yet representative set of rules with highest validity. Formally, the ***digest of rules*** is defined as a minimal subset of rules, which includes a preferable rule for any rule not in the digest. A rule A is said to be ***preferable*** to a rule B of the same class, if the rules are comparable by generality, and the rule A is more valid. The concentration algorithm finds rules in the order of their validity, with the most valid rules found first.

With both algorithms, we built only rules with maximum two conditions in their premises.

For the traditional approach, confidence threshold was selected to be 98% to exceed the proportion of the first class in data. The threshold for the support, 20%, was selected after the preliminary run of the algorithm with a lower threshold (10%) generated too many unreliable rules.

With the traditional approach and the chosen thresholds, we found 6 rules for the first class, and 72 rules for the second class (good outcome). All rules for the first class were confirmed on the test dataset, having lift more than 5.56 on the test data. The rules for the second class have lift on the

training equal 1.06; on the test, the lift ranges from 0.97 to 1.04. In all cases, the second class rules do not have practical value, since premise of any such rule improves chances of good outcome very little.

The concentration algorithm was conditioned to find not more than 75 best rules total. The algorithm found 58 rules, all of them for the class 1. All the rules had p-value by z-test below 0.001 both on the training and on the test set.

The medical application demonstrates important advantages of using a single justified validity criterion versus the traditional approach with multiple criteria.

1) One does not need to make "preliminary runs" or "guesses" to chose the thresholds for given data, given class.
2) With our approach, we were able to select the most valid and representative rules because a justified validity criterion allows to compare rules.
3) In the traditional approach, we selected rather high values for thresholds of the support and confidence. However, premises in most of the found rules have little effect on a given outcome. Contrary, all the rules selected with the z-test describe relevant conditions, significantly affecting probability of a given outcome.
4) In the traditional approach, we were able to find only six truly predictive rules. Using z-test, we found 58 such valuable rules. Many of these rules do not satisfy the high pre-set thresholds used in the traditional approach. However, the generalization ability of these dependencies is confirmed on the test set.

VI. Conclusions

We formulated requirements for a criterion of validity of probabilistic rules, to take into account all the important aspects of rules quality, such as support, confidence, and lift.

We demonstrated that the z-test of difference of proportions satisfies all the introduced axioms. We suggested to use this criterion to filter out statistically insignificant rules as well.

On the real life medical data, we showed that the z-test can replace multiple criteria, used in most studies to select rules. Having a single validity function streamlines discovery of the "interesting" rules, makes rules comparison more objective and comprehensive.

References

[1] Agrawal R, Mannila H, Strikant R, Toivonen H and Verkamo AI (1995): Fast discovery of association rules. Advances in knowledge discovery and data mining. AAAI/MIT Press, Cambridge, MA.

[2] Ziarko W (2005) Probabilistic Rough Sets. Rough sets, fuzzy sets, data mining, and granular computing. Lecture notes in artificial intelligence N 3641. Dominik Slezak et al (eds). Springer, Berlin.

[3] Webb G, Zhang, S (2005) K-Optimal Rule Discovery. Data Mining and Knowledge Discovery, 10, 3979.

[4] Klosgen W (1996) Multipattern and multistrategy discovery assistant. Advances in knowledge discovery and data mining. Usama Fayyad et al (eds). AAAI press/MIT press, Cambridge MA.

[5] Fleiss J.L., Levin B., Myunghee C.P. (2003) Statistical Methods for Rates and Proportions. Third edition. Wiley & Sons.

[6] Zaki M J, Hsiao C-J. Effcient Algorithms for Mining Closed Itemsets and Their Lattice Structure, IEEE Transaction on Knowledge and Data Engi- neering, Vol 17, No. 4, pp. 462-478, April 2005.

[7] Webb G, Zhang, S (2005) K-Optimal Rule Discovery. Data Mining and Knowledge Discovery, 10, 3979.

[8] Sapir M, Verbel D, Kotsianti A, Saidi O (2005) Live Logic: Method for Approximate Knowledge Discovery and Decision Making. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 10th International Conference, RSFDGrC 2005, Regina, Canada, August 31 - September 3, 2005, Proceedings, Part I. Lecture Notes in Computer Science 3641, 532-540

[9] Sapir M, Teverovskiy M (2006) Finding digest of rules: toward data-driven data mining. Computational intelligence. Editor: B. Kovalerchuk. Proceedings of the Second IASTED International Conference, 475 478.