On Information-Theoretic Measures of Attribute Importance

Y.Y. Yao, S.K.M. Wong, and C.J. Butz

Department of Computer Science, University of Regina
Regina, SK, S4S 0A2, Canada

Abstract. An attribute is deemed important in data mining if it partitions the database such that previously unknown regularities are observable. Many information-theoretic measures have been applied to quantify the importance of an attribute. In this paper, we summarize and critically analyze these measures.

1 Introduction

Watanabe [21] suggested that pattern recognition is essentially a conceptual adaptation to the empirical data in order to see a form in them. The form is interpreted as a structure which always entails small entropy values. Many of the algorithms in pattern recognition may be characterized as efforts to minimize entropy [20]. The philosophy of entropy minimization in pattern recognition can be applied to related fields, such as classification, data analysis, machine learning, and data mining, where one of the tasks is to discover patterns or regularities in a large data set. Regularities and structureness are characterized by small entropy values, whereas randomness is characterized by large entropy values.

One may partition the statistical population into smaller populations using the values taken by an attribute. Such an attribute is deemed important for data mining if regularities are observable in the smaller populations, while being unobservable in the statistical population. In other words, if an attribute is used for data mining, then the attribute should lead to entropy reduction. The well known ID3 inductive learning algorithm [16] uses exactly such a measure for attribute selection in a learning process. Based on the philosophy of entropy minimization, this paper examines information-theoretic measures [2, 18] for evaluating attribute importance in data mining.

2 Measuring Attribute Importance

Let X denote a discrete random variable and x_i a value in the domain of X. A joint probability distribution is a real-valued function P_X over X such that $0 \leq P_X(x_i) \leq 1$ and $\sum_{i=1}^{n} P_X(x_i) = 1$, where n denotes the number of elements in the domain of X. We write P_X as P if X is understood. Shannon's entropy function H is defined over P as:

$$H(P) = -\sum_{i=1}^{n} P(x_i) \log P(x_i),$$

where $P(x_i) \log P(x_i) = 0$ if $P(x_i) = 0$. We say Shannon's entropy is over X and write $H(P)$ as $H(X)$ when the distribution P over X is understood. Shannon's entropy is a nonnegative function, i.e., $H(X) \geq 0$. It reaches the maximum value $\log n$ when P is the *uniform* distribution, i.e., $P(x_1) = \ldots = P(x_n) = \frac{1}{n}$. The minimum entropy value 0 is obtained when the distribution P focuses on a particular value x_j, i.e., $P(x_j) = 1$ and $P(x_i) = 0,\ 1 \leq i \leq n,\ i \neq j$.

The conditional entropy, i.e., the difference between joint entropy and marginal entropy, is given by:

$$H(X \mid Y) = H(X, Y) - H(Y).$$

Mutual information can be defined as:

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = H(X) + H(Y) - H(X, Y).$$

That is, the mutual information measures the decrease of uncertainty about X caused by the knowledge of Y, and vice versa. It is a measure of the amount of information about X contained in Y. This measure is the same as the amount of information about Y contained in X, namely, $I(X;Y) = I(Y;X)$. Furthermore, the amount of information contained in X about itself is obviously $H(X)$, namely, $I(X;X) = H(X)$.

One may view an attribute and a database as a statistical variable taking values from its domain and a statistical population, respectively [5, 14]. Information-theoretic measures quantify relationships between random variables. They can immediately be applied for the analysis of databases and the evaluation of the usefulness of attributes in data mining [14].

One of the main tasks in knowledge discovery and data mining (KDD) is to find important relationships, or associations, between attributes. In statistical terms, two attributes are associated if they are not independent [11]. Two attributes are independent if changing the value of one does not affect the value of the other. From this standpoint, we comment on the meaning of information-theoretic measures in the context of data mining.

For an attribute (or a set of attributes) X, the entropy value $H(X)$ indicates the information uncertainty associated with X. An attribute with a very large domain normally divides the database into more smaller classes than an attribute with a small domain. A regularity found in a very small portion of database may not necessarily be useful. On the other hand, an attribute with small domain usually divides the database into a few larger classes. One may not find regularities in such large subsets of the database. Entropy values may be used to control the selection of attributes. It is expected that an attribute with middle range entropy values may be more useful. Similar ideas have been used successfully in information retrieval [22]. A high frequency term tends to have a higher entropy value, and a lower frequency term tends to have a lower entropy value. Both may not be good index terms. The middle frequency terms tend to be useful in describing documents in a collection.

The conditional entropy $H(Y|X)$ measures the degree of one-way implication or functional dependency of the sets of attributes X and Y. If the functional dependency $X \rightarrow Y$ holds, we conclude that $P(y_j|x_i)$ is either 1 or 0. In

term of conditional entropy, $X \rightarrow Y$ holds if and only if $H(Y|X) = 0$ [10, 14]. By the relationships between entropy, conditional entropy, and mutual information, the above condition can be equivalently stated as $H(X) = H(X,Y)$ or $I(X;Y) = H(Y)$ [10]. If Y is dependent on X, the partition of the database by X and Y is exactly the same as the one produced by X alone. The former condition reflects this observation. The latter condition shows that the mutual information between X and Y is the same as the self-information of Y. The conditional entropy function can be used to measure the importance of attributes for discovering one-way associations. For a fixed Y, one obvious disadvantage of using $H(Y|X)$ is that it favours attributes with large domains, namely, attributes with high entropy values [16].

Mutual information measures the degree of deviation of a joint distribution from the independence distribution [21]. It may be used to evaluate the usefulness of attributes in finding two-way associations. With a fixed Y, the use of $I(X;Y)$ for finding a two-way association is in fact the same as using $H(Y|X)$ for finding a one-way association [13, 19]. Two sets of attributes X and Y are statistically independent if $I(X;Y) = 0$. Equivalently, we can state this condition as $H(X) = H(X|Y)$, $H(Y) = H(Y|X)$, or $H(X,Y) = H(X) + H(Y)$. If X and Y are independent, one cannot use values of X to predicate the values of Y, and vice versa. In information-theoretic terms, knowing the value of Y does not reduce our uncertainty about X, and vice versa.

Conditional entropy and mutual information serve as the basic quantities for measuring attribute associations. By combination and normalization, one may obtain many information-theoretic measures of attribute importance. In summary, the following three groups can be obtained:

- Lee [10], Malvestuto [14], Pawlak *et al.* [15]: $H(X \mid Y), H(Y \mid X)$;
  Kvålseth [9], Malvestuto [14], Quinlan [16]: $I(X;Y)/H(X), I(X;Y)/H(Y)$.
- Knobbe and Adriaans [8], Linfoot [12], Quinlan [16]: $I(X;Y)$;
  Malvestuto [14]: $I(X;Y)/H(X,Y)$; Kvålseth [9]: $2I(X;Y)/(H(X) + H(Y))$;
  Horibe [4], Kvålseth [9]: $I(X;Y)/\max(H(X), H(Y))$;
  Kvålseth [9]: $I(X;Y)/\min(H(X), H(Y))$.
- López de Mántaras [13], Wan and Wong [19]: $H(X \mid Y) + H(Y \mid X)$;
  López de Mántaras [13], Rajski [17]: $(H(X \mid Y) + H(Y \mid X))/H(X,Y)$.

Measures in the first group are asymmetric while measures in the other two groups are symmetric. Measures in the third group are distance measures. One can obtain the following relationships between these measures:

$$\text{(i)} \quad \frac{I(X;Y)}{H(X)} = 1 - \frac{H(X|Y)}{H(X)},$$

$$\text{(ii)} \quad \frac{I(X;Y)}{\max(H(X), H(Y))} = \min\left(\frac{I(X;Y)}{H(X)}, \frac{I(X;Y)}{H(Y)}\right),$$

$$\text{(ii)} \quad \frac{I(X;Y)}{\min(H(X), H(Y))} = \max\left(\frac{I(X;Y)}{H(X)}, \frac{I(X;Y)}{H(Y)}\right),$$

$$\text{(iv)} \quad 0 \leq \frac{I(X;Y)}{\max(H(X),H(Y))} \leq \frac{2I(X;Y)}{H(X)+H(Y)} \leq \frac{I(X;Y)}{\min(H(X),H(Y))},$$

$$\text{(v)} \quad H(X|Y)+H(Y|X) = H(X,Y) - I(X;Y),$$

$$\text{(vi)} \quad \frac{2I(X;Y)}{H(X)+H(Y)} = 2\left(1 - \frac{H(X,Y)}{H(X)+H(Y)}\right),$$

$$\text{(vii)} \quad \frac{H(X \mid Y)+H(Y \mid X)}{H(X,Y)} = 1 - \frac{I(X;Y)}{H(X,Y)}.$$

They provide additional support for various measures. Furthermore, measures of one-way association can be expressed in a general form as different normalizations of conditional entropy, while measures of two-way association as different normalizations of mutual information [9].

In studying main problems for KDD, Klösgen [7] discussed two types of problems, namely, *classification and predication* and *summary and description*. Kamber and Shinghal [6] referred to them as the discovery of discriminant and characteristic rules, respectively. The classification and predication problem deals with the discovery of a set of rules or similar patterns for predicting the values of a dependent variable. The ID3 algorithm [16] and the mining of associate rules [1] are examples for solving this type of problem. The summary and description problem involves the discovery of a dominant structure that derives a dependency. It is important to note that asymmetric measures may be suitable for former problem, while symmetric measures may be appropriate for the latter.

In the study of association of random variables using statistical measures, Liebetrau [11] pointed out that many symmetric measures do not tell us anything about causality. When two attributes are shown to be correlated, it is very tempting to infer a cause-and-effect relationship between them. It is very important to realize that the mere identification of association does not provide grounds to establish causality. Garner and McGill [3] showed that information analysis is very similar to analysis of variance. One may then extend the argument of Liebetrau [11] to information-theoretic measures. In order to establish causality, we need additional techniques in data mining.

3 Conclusion

This preliminary study has demonstrated that asymmetric measures quantify one-way association and are typically related to conditional entropy, while symmetric measures quantify two-way association and are typically related to mutual information. If information theory is to be used to develop a formal theory for knowledge discovery and data mining, then the principle of entropy reduction and models in which causality can be established [11] warrant more attention.

References

1. Agrawal, R., Imielinski, T., and Swami, A.: Mining association rules between sets of items in large databases. Proc. of the ACM SIGMOD International Conference on the Management of Data (1993) 207-216.

2. Cover, T. and Thomas, J.: Elements of Information Theory. John Wiley & Sons, Toronto (1991)
3. Garner, W.R. and McGill, W.J.: Relation between information and variance analyses. Psychometrika **21** (1956) 219-228.
4. Horibe, Y.: Entropy and correlation. IEEE Trans. Syst. Man Cybern. **SMC-15** (1985) 641-642.
5. Hou, W.: Extraction and applications of statistical relationships in relational databases. IEEE Trans. Knowl. Data Eng. **8** (1996) 939-945.
6. Kamber, M. and Shinghal, R.: Evaluating the interestingness of characteristic rules. Proc. of KDD-96 (1996) 263-266.
7. Klösgen, W.: Explora: a multipattern and multistrategy discovery assistant. in: Fayyad, U.M, Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI Press / MIT Press, California. (1996) 249-271.
8. Knobbe, A.J. and Adriaans P.W.: Analysis of binary association. Proc. of KDD-96 (1996) 311-314.
9. Kvålseth, T.O.: Entropy and correlation: some comments. IEEE Trans. Syst. Man Cybern. **SMC-17** (1987) 517-519.
10. Lee, T.T.: An information-theoretic analysis of relational databases – part I: data dependencies and information metric. IEEE Trans. Soft. Eng. **SE-13** (1987) 1049-1061.
11. Liebetrau, A.M.: Measures of Association. Sage Publications, Beverly Hills. (1983).
12. Linfoot, E.H.: An informational measure of correlation. Information and Control **1** (1957) 85-87.
13. López de Mántaras, R.: ID3 revisited: a distance-based criterion for attribute selection. in: Ras, Z.W. (Ed.), Methodologies for Intelligent Systems, 4. North-Holland, New York. (1989) 342-350.
14. Malvestuto, F.M.: Statistical treatment of the information content of a database. Inf. Syst. **11** (1986) 211-223.
15. Pawlak, Z., Wong, S.K.M., and Ziarko, W.: Rough sets: probabilistic versus deterministic approach. Int. J. Man-Machine Stud. **29** (1988) 81-95.
16. Quinlan, J.R.: Induction of decision trees. Machine Learning **1** (1986) 81-106.
17. Rajski, C.: A metric space of discrete probability distributions. Information and Control **4** (1961) 373-377.
18. Sheridan, T.B. and Ferrell, W.R.: Man-Machine Systems: Information, Control, and Decision Models of Human Performance. MIT Press, Cambridge. (1974).
19. Wan, S.J. and Wong, S.K.M.: A measure for attribute dissimilarity and its applications in machine learning. in: Janicki, R. and Koczkodaj, W.W. (Eds.), Computing and Information. North-Holland, Amsterdam. (1989) 267-273.
20. Wang, Q.R. and Suen, C.Y.: Analysis and design of a decision tree based on entropy reduction and its application to large character set recognition. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-6** (1984) 406-417.
21. Watanabe, S.: Pattern recognition as a quest for minimum entropy. Pattern Recognit. **13** (1981) 381-387.
22. Wong, S.K.M. and Yao, Y.Y.: A probability distribution model for information retrieval. Inf. Process. Manage. **25** (1989) 39-53.

This article was processed using the LaTeX macro package with LLNCS style