BiAS: A Theme Metric to Model Mutual Association

Ramkishore Bhattacharyya

Microsoft Corporation, Redmond, WA - 98052
ramb@microsoft.com

Abstract. Identifying likeness between events is one of the fundamental necessities in machine learning and data mining techniques. Though grouping of events usually happens on their proximity in Euclidean space or the degree of similarity or the extent of linear dependence, certain applications like keyword and document clustering, phylogenetic profiling and feature selection tend to yield better results if events are grouped based on their mutual association. This paper presents a metric, the Bidirectional Association Similarity (*BiAS*) to quantify the degree of mutual association between a pair of events. We put forward generalized formulation to compute *BiAS* and establish unidirectional correspondence with the Jaccard and the cosine similarities. The measure can be suitably incorporated with clustering algorithms in grouping mutually associative events with adding precision to the discovered knowledge.

Keywords: Bi-directional association similarity, BiAS, clustering, cosine similarity, Jaccard index, mutual association.

1 Introduction

Identifying similarity or dissimilarity between events is one of the fundamental necessities in machine learning, data mining techniques which is measurable through computation of certain statistical metrics e.g. the Euclidean distance (*E*), the Jaccard (*J*) [4] [5] and cosine similarity (*C*) or the Pearson correlation coefficient (*P*). Though grouping of events, in general, happens on their proximity in Euclidean space or degree of similarity or extent of linear dependence, certain applications like document clustering, phylogenetic profiling, feature selection etc. tend to yield better results if events are grouped on the basis of pair-wise mutual association. A pair of events g_i and g_j is mutually associative if the probability of occurrence of g_j is high given the occurrence of g_i and vice-versa. Unfortunately, neither similarity nor correlation guarantees mutual associativity as they do not have individual control over $p(g_j/g_i)$ or $p(g_i/g_j)$, p denotes probability. Two events g_i and g_j can be similar or correlated if,

- there exist unidirectional association of the form $g_i \rightarrow g_j$ or
- there really exist mutual association of the form $g_i \rightarrow g_j$ and $g_j \rightarrow g_i$ or
- there exist a third event g_k such that $g_k \rightarrow g_i$ and $g_k \rightarrow g_j$ hold simultaneously

So existence and directionality of association remain vague unless explicitly verified.

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 67–73, 2013.

The *Bidirectional Association Similarity* (*BiAS*) models mutual associativity with the help of two simultaneous associations $g_i \rightarrow g_j$ and $g_j \rightarrow g_i$ [1] and is quantified by $\beta(g_i,g_j) = p(g_j/g_i)*p(g_i/g_j)$. Given two pre-specified thresholds μ and τ, g_i and g_j are mutually associative if both of $p(g_j/g_i)$, $p(g_i/g_j) \geq \mu$ and $\beta(g_i,g_j) \geq \tau$. We prove subsequently that mutually associative events are also similar, but similarity does not guarantee mutual associativity. Hence, *BiAS* adds both way associativity constraints on top of similarity. It helps in pruning out loosely coupled expression vectors with adding precision to discovered knowledge. The salient contributions in this paper are:

- Concept and foundation of *BiAS* on the basis of mutual associativity
- Formulation of *J* as a function of $p(g_j/g_i)$, $p(g_i/g_j)$ and deriving its lower bound
- Generalized formulation of *BiAS* for real-valued attributes
- Bridging connections between the lower bounds of *BiAS*, *J* and *cos*
- Finally, proving capability of *BiAS* to be integrated with clustering algorithms

2 Mutual Association and Generalized Jaccard Index

Let us assume $G = \{g_1, g_2, \ldots, g_n\}$ be a set of n events where each of g_i is a d-dimensional Boolean vector. For any g_i, a '1' (or '0') at l^{th} dimension indicates its presence (or absence) in l^{th} experiment. Further, let $T(g_i)$ be a set of integers $j \in \{1, 2, \ldots, d\}$, so that $g_{ij} = 1$. Hence, $T(g_i) \cap T(g_j)$ is a set of integers m, where both $g_{im} = 1$ and $g_{jm} = 1$. Also, assume that $c(g_i) = |T(g_i)|$, the frequency of occurrence of the event g_i in the dataset. For preciseness, we denote $p(g_j/g_i)$ as μ_f and $p(g_i/g_j)$ as μ_b. The generalized form of *Jaccard* index *J* for a set of events G' can be formulated as:

$$J(G') = \left|\bigcap_{g \in G'} T(g)\right| \Big/ \left|\bigcup_{g \in G'} T(g)\right| .$$

Eqn. (1) in the following states a key relationship between *J*, μ_f and μ_b. It can be easily proven by replacing μ_f with $|T(g_i) \cap T(g_j)| / |T(g_i)|$ and μ_b with $|T(g_i) \cap T(g_j)| / |T(g_j)|$.

$$J(g_i, g_j) = \frac{1}{(1/\mu_f) + (1/\mu_b) - 1} \tag{1}$$

Given μ_f, μ_b and the constraints, eqn. (1) can be utilized as the criterion function to be minimized through non-linear programming optimization to yield a lower bound on *J*.

Lemma 1. Let $\mu_f, \mu_b \geq \mu$ and $\mu_f * \mu_b \geq \tau$. The *Jaccard* index of two events g_i and g_j,

$$J(g_i, g_j) \geq \frac{\mu\tau}{\mu^2 + \tau - \mu\tau} \tag{2}$$

Proof. The inequality is proven via the Karush-Kuhn-Tucker (KKT) theorem.

The Karush-Kuhn-Tucker (KKT) theorem. If the function f(x) has a minimum at x* in the feasible set and if $\nabla f(x^*)$ and $\nabla g_i(x^*)$, i = 1,2,…,m, exist ('∇' denotes partial derivative with respect to all x_i), then there exist an m-dimensional vector λ such that

- $\nabla f(x^*) + \Sigma_{i=1}^{m} \lambda_i \nabla g_i(x^*) = 0$, for i = 1, 2,…, m.
- $g_i(x^*) \bullet 0$, for i = 1, 2,…, m.

- $\lambda_i [g_i(x^*) - b_i] = 0$, for i = 1, 2,…, m.
- $\lambda_i \geq 0$, for i = 1, 2,…, m.

(x*, λ) is called a KKT point, λ is the Dual Vector or the Lagrange Multiplier.

Minimizing eqn. (1) is equivalent to maximizing $(1/\mu_f) + (1/\mu_b) - 1$ which is again equivalent to minimizing $1 - (1/\mu_f) - (1/\mu_b)$. Let us now formulate the problem: Minimize: $f(\mu_f, \mu_b)$: $1 - (1/\mu_f) - (1/\mu_b)$…(i), subjected to: g_1: $\mu_f \bullet \mu \Rightarrow -\mu_f \bullet -\mu$, g_2: $\mu_b \bullet \mu \Rightarrow -\mu_b \bullet -\mu$ and g_3: $\mu_f * \mu_b \bullet \tau \Rightarrow -\mu_f * \mu_b \bullet -\tau$. Clearly, $\nabla f(\mu_f, \mu_b)$ and $\nabla g_i(\mu_f, \mu_b)$, for i = 1, 2, 3 exist. Hence, by KKT theorem, $(1/\mu_f^2) - \lambda_1 - \lambda_3 * \mu_b = 0$ … (ii) and $(1/\mu_b^2) - \lambda_2 - \lambda_3 * \mu_f = 0$ … (iii). Also, $-\lambda_1(\mu_f - \mu) = 0$… (iv), $-\lambda_2(\mu_b - \mu) = 0$… (v), $-\lambda_3(\mu_f * \mu_b - \tau) = 0$… (vi) and $\lambda_i \bullet 0$, for i = 1, 2, 3… (vii).

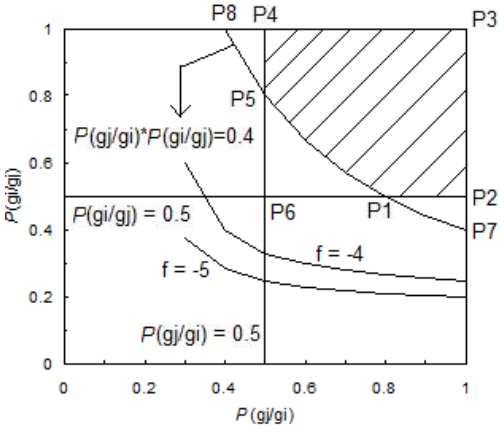

Fig. 1. Plot for eqns. μ_f=0.5, μ_b=0.5, $\mu_f * \mu_b$=0.4 and f, shaded area shows solution region

Without any constraint on μ_f and μ_b, $0 \bullet \mu_f, \mu_b \bullet 1$. Different eqns. viz. $\mu_f = p(g_j/g_i) = \mu$, $\mu_b = p(g_i/g_j) = \mu$ and $\mu_f * \mu_b = p(g_j/g_i) * p(g_i/g_j) = \tau$ (for $\mu = 0.5$ and $\tau = 0.4$) are plotted in Fig. 1. The objective function (eqn. (i)) is also plotted for f = -4 and f = -5. As can be noticed, f is minimized in the direction toward the center (0,0). Also, the feasible region for μ_f and μ_b is shaded in Fig. 1. Clearly, this is a convex region as all the points on the line segment joining any two points from the region completely remain within it. The regions inside P1P2P7 and P4P5P8 get excluded from the solution due to the constraint on $p(g_j/g_i)$ or $p(g_i/g_j)$. The constraint on $p(g_j/g_i) * p(g_i/g_j)$ excludes the region P1P5P6 from the solution. Together, they ensure high confidence in one direction with a minimal in the reverse. It is evident from eqn. (i) that both of μ_f and μ_b need to be minimized for minimization of f. Hence the solution must be somewhere on the line segment P1P5.

Consider Point P1. P1 is the point of intersection of $\mu_b = \mu$ and $\mu_f * \mu_b = \tau$. At P1, $\mu_f \neq \mu$. Hence, from eqn. (iv), $\lambda_1 = 0$. Given this, eqn. (ii) yields $\lambda_3 = 1/(\mu_f^2.\mu_b)$. Replacing λ_3 with its value in eqn. (iii), $\lambda_2 = (\mu_f - \mu_b) / (\mu_f.\mu_b^2) = (\tau - \mu^2) / (\mu_f^2.\mu^2)$ (replacing μ_b by μ and μ_f in the numerator by τ/μ_b). Since, $\tau \bullet \mu^2$, $(\tau - \mu^2) \bullet 0$. **Thus, all** λ_1, λ_2 and λ_3 are

• 0, which satisfies eqn. (vii). Putting $\mu_b = \mu$ and $\mu_f = \tau / \mu$ in the objective function (in (i)) f = 1 – (μ / τ) – (1 / μ) … (viii).

Consider Point P5. Similarly, on point P5, it can be shown that all λ_1, λ_2 and λ_3 are • 0 and value of the objective function (in (i)) f = 1 – (μ / τ) – (1 / μ), same as (viii).

Consider Any Point on the Line Segment P1P5. Here, μ_f, $\mu_b \neq \mu$ but $\mu_f * \mu_b = \tau$. From eqn. (iv), (v) and (vi), $\lambda_1 = 0$ and $\lambda_2 = 0$. Putting this in eqn. (ii) and (iii), $\lambda_3 = 1/\mu_f^2.\mu_b = 1/\mu_f.\mu_b^2 \Rightarrow \mu_f = \mu_b = \tau^{0.5}$. Thus, all λ_1, λ_2 and λ_3 are • 0, which satisfies eqn. (vii) and f = 1 – (1 / $\tau^{0.5}$) – (1 / $\tau^{0.5}$) = 1 – (2 / $\tau^{0.5}$)… (ix).
(viii) – (ix) = (2/$\tau^{0.5}$) – (μ/τ) – (1/μ) = – (τ + μ^2 – 2 . $\mu\tau^{0.5}$) / $\mu\tau$ = – ($\tau^{0.5}$ – μ)^2 / $\mu\tau$ • 0. Hence, (viii) • (ix) and optimum solution exist at P1 or P5. For P1, $\mu_f = \tau/\mu$ and $\mu_b = \mu$, for P5 $\mu_f = \mu$ and $\mu_b = \tau/\mu$. Putting, the values in eqn. (4), $\xi(g_i,g_j)$ • 1 / ((μ/τ) + (1/μ) – 1) = $\mu\tau$ / (μ^2 + τ – $\mu\tau$). Thus eqn. (4) follows. □

3 Generalization of *BiAS* for Real-Valued Attributes

We now waive the restriction to Boolean attributes and assume $g_i = [g_{i1}, g_{i2}, \ldots, g_{id}]$, $g_{ik} \in \Re$, the set of real. With that, $c(g_i)$ is redefined to square of the L_2-norm of g_i:

$$c(g_i) = \|g_i\|^2 = \sum_{k=1}^{d} g_{ik}^2 \tag{3}$$

The frequency of joint occurrence $c(g_i \cap g_j)$ is formulated by their dot product:

$$c(g_i \cap g_j) = g_i \bullet g_j = \sum_{k=1}^{d} g_{ik} * g_{jk} \tag{4}$$

With eqn. (3) and (4), computation of $p(g_j/g_i)$, $p(g_i/g_j)$ is immediate. Also, $J(g_i,g_j) = c(g_i \cap g_j) / c(g_i \cup g_j) = c(g_i \cap g_j) / c(g_i) + c(g_j) - c(g_i \cap g_j)$. Thus, the Jaccard index (more specifically the Tanimoto coefficient [7]) and *BiAS* can be reformulated as:

$$J(g_i, g_j) = \frac{g_i \bullet g_j}{\|g_i\|^2 + \|g_j\|^2 - g_i \bullet g_j} \tag{5}$$

$$\beta(g_i, g_j) = p(g_j / g_i) * p(g_i / g_j) = \frac{[g_i \bullet g_j]^2}{\|g_i\|^2 * \|g_j\|^2} = \frac{(\sum_{k=1}^{d} g_{ik} * g_{jk})^2}{\sum_{k=1}^{d} g_{ik}^2 * \sum_{k=1}^{d} g_{jk}^2} \tag{6}$$

The physical significance of probability of a real-valued attribute may not be as straight-forward as it is for Boolean attributes. So, eqns. (5) and (6) are represented with the frequency of occurrences of an event which holds perfect without any loss of generality. Under the new formulations, $p(g_j/g_i)$, $p(g_i/g_j)$ can be more than 1 or even negative (signifying opposite similarity). Nevertheless, the absolute value of their product remains within [0,1]. Lemma (2) proves it.

Lemma 2. For any two *d*-dimensional real-valued vectors g_i and g_j,

$$0 \leq \beta(g_i, g_j) \leq 1. \tag{7}$$

Proof. Rewriting eqn. (6),

$\beta(g_i, g_j) = \frac{[g_i \bullet g_j]^2}{\|g_i\|^2 * \|g_j\|^2} = \left[\frac{g_i}{\|g_i\|} \bullet \frac{g_j}{\|g_j\|}\right]^2$. So $\beta(g_i,g_j)$ is the square of the scalar product of two unit vectors. Hence, $\beta(g_i,g_j) \geq 0$ and $\beta(g_i,g_j) \leq 1$. □

4 Connecting Lower Bounds of Different Metrics

Sometimes, a particular metric can be suitable for certain applications but the choice of threshold may not be apparent due to the physical significance not being straightforward in context of those applications. Connecting the lower bounds of different similarity metrics, particularly with different physical interpretations like similarity or mutual dependence, may help in this regard. Note that we assume $p(g_j/g_i), p(g_i/g_j) \geq \mu$, unless otherwise specified.

Relation between the lower bounds of $J(g_i,g_j)$ and $\beta(g_i,g_j)$ is derived in eqn. (2) which involves the threshold μ as well. For example, with $\mu = 0.5$ and $\tau = 0.37$ (the average of μ and μ^2), lower bound of Jaccard similarity is (0.5*0.37) / (0.5^2 + 0.37 − 0.5*0.37) = 0.185 / 0.435 = 0.425. In case τ is equal to its lower bound $(0.5)^2 = 0.25$, lower bound of $J(g_i,g_j)$ is (0.5*0.25) / (0.5^2 + 0.25 − 0.5*0.25) = 0.125 / 0.375 = 0.33. The choice of τ helps to raise the threshold of J from 0.33 to 0.425 and thus pruning additional pairs with similarities in between those two values.

From eqn. (6),

$\beta(g_i, g_j) = \frac{[g_i \bullet g_j]^2}{\|g_i\|^2 * \|g_j\|^2} = \cos^2(g_i, g_j)$. So, $cos(g_i,g_j) \geq \beta(g_i,g_j)$ as $0 \leq \beta(g_i,g_j) \leq 1$.

This establishes that mutually associative events are also similar.

Finally, we deduce the relation between lower bounds of the $J(g_i,g_j)$ and $cos(g_i,g_j)$. As $\beta(g_i,g_j) \geq \mu$, $|cos(g_i,g_j)| \geq \sqrt{\tau}$ as cosine similarity can be negative as well. Let us denote the lower bound of $|cos(g_i,g_j)|$ as $lb(|cos(g_i,g_j)|)$ and assume that $\tau = \mu^2$, where $p(g_j/g_i), p(g_i/g_j) \geq \mu$. Replacing this in eqn. (2),

$$J(g_i, g_j) \geq \frac{\mu\tau}{\mu^2 + \tau - \mu\tau} = \frac{\tau^{3/2}}{2\tau - \tau^{3/2}} = \frac{\tau^{3/2}}{\tau(2 - \sqrt{\tau})} = \frac{\sqrt{\tau}}{(2 - \sqrt{\tau})} = \frac{lb(|\cos(g_i, g_j)|)}{2 - lb(|\cos(g_i, g_j)|)}.$$

So, given the cosine similarity of g_i and g_j we can get the lower bound of $J(g_i,g_j)$.

5 Clustering Using *BiAS*

Clustering is one of the most popular and well-established unsupervised data mining techniques that deal with finding a structure in a collection of unlabeled data and determining the intrinsic grouping. The *k*-means clustering algorithm [6] is perhaps the most popular iterative solution to group events in a pre-specified *k* number of clusters. The idea is to gather all those events which lie within the preset similarity or distance from the cluster center. Thus, to ensure that *BiAS* can successfully identify a set of mutually associative events, we must prove the following two properties:

- If two events g_i and g_j happen to be mutually associative individually with a third event g_c, g_i and g_j are also mutually associative with respect to certain threshold
- If there exist pair-wise mutual association between any two events of g_i, g_j and g_c, *Jaccard* index of g_i, g_j and g_c together has a lower bound.

Lemma 3 and 4 state the lower bounds, the proofs are beyond the scope of this paper.

Lemma 3. Let g_i, g_j and g_c are three events such that $p(g_i/g_c) = \mu_1$, $p(g_c/g_i) = \mu_2$, $p(g_j/g_c) = \mu_3$ and $p(g_c/g_j) = \mu_4$. Then,

$$p(g_j / g_i) \geq \mu_2.(\mu_1 + \mu_3 - 1)/\mu_1 \tag{8}$$

$$p(g_i / g_j) \geq \mu_4.(\mu_1 + \mu_3 - 1)/\mu_3 \tag{9}$$

For example, if $\mu = 0.85$ and μ_1, μ_2, μ_3, $\mu_4 \geq \mu$, $p(g_j/g_i)$ and $p(g_i/g_j)$ will be greater than $\mu.(\mu + \mu - 1) / \mu = \mu\,(2\mu - 1) = 0.85*0.7 = 0.6$. So, if the events are clustered based on their mutual associativity with the cluster center, any pair of events within the cluster has a lower bound on their *BiAS* metric and hence the *J*. If so, lemma 4 states another lower bound for $J(g_i,g_j,g_c)$.

Lemma 4. Let g_i, g_j and g_c be three events and all of $J(g_i, g_c)$, $J(g_j, g_c)$ and $J(g_i, g_j)$ are greater than some threshold q. Then,

$$J(g_i, g_j, g_c) \geq 1 - 3(1-q)(2-q^2)/2. \tag{10}$$

E.g. with $q = 0.8$, $J(g_i,g_j,g_c) \geq 1 - 3*(1-0.8)*(2-0.64)/2 = 1 - 0.41 = 0.59$.

Thus, if a clustering algorithm ensures that an incoming event is mutually associative with the center, it is guaranteed to have mutual associativity as well as similarity with all other existing events which, in turn, maintain the overall cluster quality.

6 Conclusion

The *Bidirectional Association Similarity* ensures that similarity between two events results from true inter-dependence which we model through mutual conditional probabilities. Conventional similarity metrics, in general, just quantize the likeness between two expression vectors. It is not designed to capture similar expressions owing to mutual association between a pair of events. Apparently, *BiAS* is a stricter criterion ensuring both mutual association as well as similarity. This paper builds strong foundation of the measure, bridges connection with other well-known similarity metrics and theoretically proves the effectiveness in knowledge discovery.

We are currently working on a few promising application domains where mutual associativity is much apparent in natural phenomena and hence *BiAS* can be instrumental in knowledge discovery. Feature selection is definitely one of our interests where few other literatures [2] [3] have envisaged the effectiveness of mutual association. Also, identifying self-regulatory systems in genetics through feedback loop is another potential area to invest on.

References

1. Agrawal, R., Imielinski, T., Swamy, A.: Mining Association Rules between Sets of Items in Large Databases. In: Proc. ACM Int'l Conf. on Management of Data (SIGMOD), Washington, DC, pp. 207–216 (May 1993)
2. Blake, C., Pratt, W.: Better rules, fewer features: a semantic approach to selecting features from text. In: Proc. IEEE Intl. Conf. on Data Mining (ICDM 2001), pp. 59–66 (2001)
3. Dai, X., Jia, J., Ghaoui, L.E., Yu, B.: SBA-term: Sparse Bilingual Association for Terms. In: Fifth Intl. Conf. on Semantic Computing (ICSC 2011), pp. 189–192 (2011)
4. Everitt, B.: Cluster analysis, 3rd edn. Edward Arnold, London (1993)
5. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Englewood Cliffs, N.J (1998)
6. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1965)
7. Tanimoto, T.T.: IBM Internal Report November 17 (1957)