

On Mining Statistically Significant Attribute Association Information

Pritam Chanda * Jianmei Yang † Aidong Zhang ‡ Murali Ramanathan§

Abstract

Knowledge of the association information between the attributes in a data set provides insight into the underlying structure of the data and explains the relationships (independence, synergy, redundancy) between the attributes. Complex models learnt computationally from the data are more interpretable to a human analyst when such interdependencies are known. In this paper, we focus on mining two types of association information among the attributes - correlation information and interaction information which capture multivariate dependencies between the data attributes. Identifying the statistically significant attribute associations is a computationally challenging task - the number of possible associations increases exponentially and many associations contain redundant information when a number of correlated attributes are present. In this paper, we explore efficient data mining methods to discover non-redundant attribute sets that contain significant association information indicating the presence of informative patterns in the data.

Keywords: Information theory, Entropy, Attribute Association, Correlation, Interaction.

1 Introduction

Large volumes of data are being generated in various fields of scientific research, economics, financial and marketing applications. Data mining techniques have been employed to make sense of these data sets, to discover useful patterns and models in the data that aid explaining how the system being represented works. To discover key patterns in the data, it is necessary to find associations between the attributes in the data that represent the interdependencies (such as independence, synergy and redundancy) and discover their statistical properties. These relationships are also useful for understanding an appropriate probabilistic model representing the data. Often the models constructed using

some prior distributions and assumptions and the observed data become too cumbersome and lose clarity; it can yield excellent classification accuracy, but we do not really learn anything from the data for practical applications. In such cases, exploring attribute association patterns enable deeper insight into the data, and possibly allow one to comprehend the model(s) computationally learnt from the data.

From an information theoretic perspective, association information between attributes can be broadly categorized into (1) correlation information and (2) interaction information. The correlation information of an attribute set represents the total amount of information shared among the attributes; equivalently, it can be viewed as a general measure of dependency. The interaction information of an attribute set captures the multivariate dependencies between the attributes such that the information is not present in any subset of the given set. As we shall demonstrate in this paper, these two are related and complement each other in discovering useful patterns and relationships in the data.

In this paper, we study the problem of mining the above two types of association information in discrete data. Finding these associations have important implications in a biological or genetic context. The risk of developing many common and complex diseases such as cancer, autoimmune disease and cardiovascular disease involves complex interactions between multiple genes and several endogenous and exogenous environmental factors. For many complex diseases, individually each gene (or equivalently the single nucleotide polymorphisms on that gene) have weak associations with the disease, however, together they interact in a complicated fashion (often with environmental factors e.g. smoking habit) to control the expression of the disease [9, 8]. The successful detection of such gene-gene interactions and gene-environment interactions can provide the scientific basis for many underlying biological interactions, improves the prospects for uncovering potentially undiscovered genes involved in the disease process and helps to develop preventative and curative measures for particular genetic susceptibilities.

Besides genetics, the usefulness of exploring association information is also important in supervised learning problems such as feature selection where the task is to

*Dept. of Computer Science and Engineering, State University of New York, Buffalo, NY. Email:pchanda@buffalo.edu

†Dept. of Computer Science and Engineering, State University of New York, Buffalo, NY. Email:jy38@buffalo.edu

‡Dept. of Computer Science and Engineering, State University of New York, Buffalo, NY. Email:azhang@buffalo.edu

§Dept. of Pharmaceutical Sciences, State University of New York, Buffalo, NY. Email:murali@buffalo.edu

find a subset of the features that improve the accuracy of a classifier. Correlation measures such as Pearson’s correlation, Spearman’s rank correlation, Kendall tau correlation and chi square measures are commonly used to evaluate individual attribute relevancy in predicting the class attribute. Associations among attributes have been used for feature selection directly or indirectly in various data mining and machine learning applications, however most of these consider only pairwise associations (i.e. mutual information) between each attribute and the class [12, 29, 25].

Contributions of the paper: In this paper, we study the problem of mining statistically significant correlation information and interaction information in discrete data. Exploring all subsets of attributes for significant association information becomes computationally intractable as the number of attributes increases. To tackle the problem efficiently and effectively, we do the following:

1. We demonstrate and prove the relationship between the two association information metrics.
2. We derive the distributional properties of correlation information.
3. We propose the concepts of attribute combinations containing highly significant, moderately significant and non-significant correlation information. These are used to formulate *combinations of interest* as highly significant attribute sets that have all subsets with non-significant correlation information, and *special combinations of interest* that can have at most one subset with highly significant correlation information.
4. We present some bounds on correlation information and develop several pruning strategies utilizing these bounds to efficiently prune the search space.
5. Using the bounds and pruning strategies, we develop the algorithms correlation information miner (CIM) and interaction information miner (IIM).
6. Using several experimental and real-life data sets, we critically examine the effectiveness and efficiency of our proposed mining algorithms.

2 Related Work

Mining correlation information in high-dimensional discrete data has attracted much research interest in recent years. Various approaches have been developed, including correlation pattern mining [21, 19, 24], feature selection [12, 29, 25], finding correlated item pairs [33], and others. Mining correlation information is also closely associated with mining frequent patterns in the data. It roots from the association rule mining problem introduced in the Apriori algorithm [2]. Since then much

work has been done on frequent pattern mining with itemsets, constrained rule mining, measuring interestingness of association rules mined and so on. Traditionally support and confidence and related measures have been used to assess the usefulness of the rules mined. Correlation pattern mining was achieved with a statistical basis in [5] where the authors have used χ^2 correlation measure between pairs of attributes. Information theory based metrics like entropy have also been used as a quality measure for sets of attributes (or items) and efficient algorithms have been proposed to mine the maximally informative k -itemsets as in [20]. Algorithms have been proposed to find low-entropy sets as in [14] where they introduced two kinds of low entropy trees and discussed their properties. In the NIFS method [32], the authors explore the problem of finding non-redundant high order correlations in binary data and propose pruning strategies by investigating the bounds of multi-information which is a generalization of pairwise mutual information. Their proposed pruning methods are based on hard thresholds which are difficult to set unless pre-determined using trial and error. We also use pruning strategies using bounds on correlation information, however, instead of hard thresholds, we employ the distributional properties of correlation information which improves the power of our method in the presence of noise in the data. Also our bounds are based on entropy inequalities and therefore not restricted to binary data. Using experimental data sets, we further show that our methods can identify attribute sets (we call them *special combinations of interest*) which are not detected in [32] and also mine interaction information among the attribute sets.

Compared with correlation information, interaction information is a more parsimonious measure of association. Interaction information between variables and attributes was researched upon in diverse areas like physics, information theory, neuroscience, game theory, law and economics. The concept was first introduced by McGill [23] as a multivariate generalization of Shannon’s mutual information [27]. Later, Han [13] gave rigorous formal definitions of the concepts of interaction while properties of positive and negative interactions appeared in [30]. In physics, Cerf [7] analyzed interaction information of three variables in quantum physics, while Matsuda [22] studied properties of interaction information (referred to as higher order mutual information functions) for general complex systems. Bell [4] defined co-information forming a partially ordered lattice in terms of the entropies and used it for dependent component analysis. More recently, Jakulin [16, 18] studied it extensively from a machine learning perspective and provided methods for visualizing interactions and inter-

preting the structure in the data.

3 Association Information Metrics

In this section, we introduce some basic notations that we shall use throughout the paper. In the rest of the paper, the term combination is also used to refer to a set of attributes. A given data set D is represented as a $m \times n$ matrix of discrete values where each row is a sample and each column is an attribute. Let $\zeta = \{A_1; A_2; \dots; A_n\}$ be the set of attributes in D . We treat A_i as a discrete random variable and $p(a_i)$ represents the probability density function of A_i . Also, the words 'combination' and 'set' are used interchangeably in the paper referring to a collection of attributes.

DEFINITION 3.1. *The uncertainty of a discrete random variable A_i is defined by Shannon's entropy [27] as,*

$$H(A_i) = - \sum_{a_i} p(a_i) \log(p(a_i))$$

DEFINITION 3.2. *The interaction information among the k attributes (k -way interaction information or KWII) in set $S = \{A_1; A_2; \dots; A_k\}$, $S \subseteq \zeta$, is the multivariate generalizations of Shannon's mutual information. It is defined as the amount of information (synergy or redundancy) that is present in the set of attributes, such that it is not present in any subset of these attributes [16]. The KWII can be written succinctly as an alternating sum of the entropies of all possible subsets τ of S using the difference operator notation of Han [13]:*

$$KWII(S) = - \sum_{\tau \subseteq S} (-1)^{|S \setminus \tau|} H(\tau)$$

The number of attributes k in a combination is called the order of the combination. It quantifies interactions by representing the information that cannot be obtained without observing all k attributes at the same time.

In the bivariate case, the KWII is always nonnegative. But in the multivariate case, KWII can be positive (indicating synergy) or negative (indicating redundancy). A value of zero indicates the absence of k -way interactions.

DEFINITION 3.3. *The Total Correlation Information (TCI) involving attributes in set $S = \{A_1; \dots; A_k\}$ is defined [13][17] as,*

$$\begin{aligned} TCI(S) &= \sum_{i=1}^k H(A_i) - H(A_1; \dots; A_k) \\ &= \sum_{a_1, \dots, a_k} p(a_1 \dots a_k) \log_2 \left(\frac{p(a_1 \dots a_k)}{p(a_1) \dots p(a_k)} \right) \end{aligned}$$

The TCI is the total amount of information shared among the attributes in the set and has the properties: (1) A TCI value that is zero indicates that the attributes are independent. (2) The maximal value of TCI occurs when one attribute is completely redundant with the others. (3) The TCI is always non-negative and increases monotonically with increasing combination size i.e., $TCI(A_1; \dots; A_k) \leq TCI(A_1; \dots; A_k; A_{k+1})$.

4 Relationship between KWII and TCI

Here, we demonstrate the relationship between the above mentioned two information theoretic metrics.

THEOREM 4.1. *Let $S \subseteq \zeta$ be a set of attributes. Then,* $H(S) = - \sum_{Z \subseteq S} KWII(Z)$.

Proof. For $Z \subseteq S$, define,

$$(4.1) \quad \Gamma(Z) = \sum_{W \subseteq Z} (-1)^{|Z \setminus W|} \log(p(W)), \text{ and,}$$

$$(4.2) \quad KWII(A_i) = \sum_{a_i} p(a_i) \log(p(a_i)) \text{ for } A_i \in S$$

Then, $E[\Gamma(Z)] = \sum_{W \subseteq Z} (-1)^{|Z \setminus W|} E[\log(f(W))]$ where E denotes expectation. But $E[\log(f(W))] = -H(W)$ so that $E[\Gamma(Z)] = - \sum_{W \subseteq Z} (-1)^{|Z \setminus W|} H(W) = KWII(Z)$. Note that (4.1) is a Mobius transform [15] of 4.1. Taking the inverse Mobius transform, we get $f(S) = \prod_{Z \subseteq S} \exp(\Gamma(Z))$. Taking log of both sides and then expectation, we get

$$E[\log(f(S))] = \sum_{Z \subseteq S} E[\Gamma(Z)] = \sum_{Z \subseteq S} KWII(Z)$$

But $E[\log(f(S))] = -H(S)$ so that,

$$H(S) = - \sum_{Z \subseteq S} KWII(Z) \quad \blacksquare$$

THEOREM 4.2. *The TCI of an attribute set S represents the sum of all KWII between two or more attributes from S , i.e., $TCI(S) = \sum_{Z \subseteq S, |Z| \geq 2} KWII(Z)$*

Proof. Using equation 4.2,

$$\begin{aligned} \sum_{Z \subseteq S} KWII(Z) &= \sum_{A_i \in S} KWII(A_i) + \sum_{Z \subseteq S, |Z| \geq 2} KWII(Z) \\ &= - \sum_{A_i \in S} H(A_i) + \sum_{Z \subseteq S, |Z| \geq 2} KWII(Z) \end{aligned}$$

Then, by theorem 4.1 and TCI definition 3.3,

$$\sum_{A_i \in S} H(A_i) + \sum_{Z \subseteq S} KWII(Z) = \sum_{Z \subseteq S, |Z| \geq 2} KWII(Z)$$

$$\text{or, } \sum_{A_i \in S} H(A_i) - H(S) = \sum_{Z \subseteq S, |Z| \geq 2} KWII(Z)$$

$$\text{or, } TCI(S) = \sum_{Z \subseteq S, |Z| \geq 2} KWII(Z) \quad \blacksquare$$

Note that for a set with two attributes, KWII and TCI are identical and represents mutual information.

	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	TCI values
S ₁	1	0	0	1	0	0	{1,2,3} = 1.0, {4,5,6} = 1.02,
S ₂	1	0	0	0	1	1	{1,2,3,4} = 1.06,
S ₃	1	0	0	1	0	0	{1,3,4} = 0.06
S ₄	0	0	1	1	2	0	KWII values
S ₅	0	0	1	0	2	1	{1,2} = 0.0, {1,3} = 0.0,
S ₆	0	0	1	1	0	0	{1,4} = 0.02, {2,3} = 0.0,
S ₇	0	1	0	1	1	1	{2,4} = 0.02, {3,4} = 0.02,
S ₈	0	1	0	0	2	1	{4,5} = 0.04, {4,6} = 0.48,
S ₉	0	1	0	0	1	1	{5,6} = 0.38, {1,2,3} = 0.0,
S ₁₀	1	1	1	1	2	0	{1,2,4} = 0.02,
S ₁₁	1	1	1	1	1	1	{2,3,4} = 0.02,
S ₁₂	1	1	1	0	0	1	{4,5,6} = 0.13

Figure 1: A Toy Example.

Consider the discrete data set shown Figure 1. Low TCI of the set $\{A_1; A_3; A_4\}$ indicates no association present. A relatively higher TCI value of 1.0 reflects the useful correlation information contained in the set $\{A_1; A_2; A_3\}$. The set $\{A_1; A_2; A_3; A_4\}$ also has a higher TCI of 1.06, but note that increase in TCI by inclusion of A_4 is only 0.06 and all proper subsets of $\{A_1; A_2; A_3; A_4\}$ containing A_4 has small KWII indicating that A_4 does not contain useful association information in combination with A_1, A_2 and A_3 . Again, KWII of $\{A_1; A_2; A_3\}$ is 1.0 which causes the TCI to be 1.0 and TCI of $\{A_4; A_5; A_6\}$ is 1.02 with contributions mainly from $KWII(A_4; A_6) = 0.48$, $KWII(A_5; A_6) = 0.38$ and $KWII(A_4; A_5; A_6) = 0.13$. These reflect that KWII represents more parsimonious chunks of association information that constitutes the TCI and reveals the structure in the data that is not obvious from the correlation information.

5 Problem Formulation

In this section, we introduce the concepts of Combinations of Interest (or COI) and Special Combinations of Interest (or SCOI). A COI is an attribute set containing high TCI such that its proper subsets have low TCI, while a SCOI is similar to the COI but can have exactly one proper subset to have high TCI. Our definitions of high and low are based on statistical significance levels which is based on distributional properties explored in section 7. Broadly, our goal is to mine the COI, SCOI and combinations with high KWII that represent attribute sets containing non-redundant association information. To develop our mining strategy, we first give some formal definitions.

5.1 Preliminary Definitions Assume that we know the probability distribution function of the TCI metric. Let α_{High} and α_{Low} are two given significance levels for determining the statistical significance of an observed TCI such that $0 < \alpha_{High} < \alpha_{Low}$. Let $S = \{A_1; \dots; A_k\} \subseteq \zeta$ be a given set of attributes.

DEFINITION 5.1. S has statistically **Highly Signif-**

icant correlation information if $Pvalue(TCI(S)) < \alpha_{High}$. We refer to such a combination of attributes as **Highly Significant Combination** or **HSC**.

DEFINITION 5.2. S has statistically **Non-Significant** correlation information if $Pvalue(TCI(S)) \geq \alpha_{Low}$. We refer to such a combination of attributes as **Non-Significant Combination** or **NSC**.

DEFINITION 5.3. S has statistically **Moderately-Significant** correlation information if $\alpha_{High} \leq Pvalue(TCI(S)) < \alpha_{Low}$. Such a combination of attributes is called a **Moderately-Significant Combination** or **MSC**.

For example, setting $\alpha_{High} = 10^{-10}$ and $\alpha_{Low} = 10^{-3}$, a $Pvalue$ of 10^{-12} will be Highly Significant while that of 0.01 will be Non-Significant.

DEFINITION 5.4. S is a **Combination Of Interest** (or **COI**) if it satisfies:-

1. S is a HSC, and
2. Each proper subset of S is a NSC.

However checking all proper 2^{k-1} subsets of S is computationally expensive. Let $S_{k-1} \subset S$ with $k-1$ attributes. From the monotonic increasing property of the TCI (property (3) in definition 3.3), $TCI(S) \geq TCI(S_{k-1})$. Therefore, we make the assumption that if $Pvalue(TCI(S)) \geq \alpha_{Low}$, then $Pvalue(TCI(S_{k-1}))$ is also $\geq \alpha_{Low}$ as smaller TCI value usually has lower significance. As a result, we only need to check whether the $k-1$ size subsets of S are NSC.

The definition of COI is based on the fact that if S is a HSC and one or more of its subsets are HSC or MSC, then S has redundancy as it has at least one subset with high correlation information. For example, assume set $S = \{A_1; A_2; A_3; A_4\}$ is a HSC and its subsets $S' = \{A_1; A_2\}$ and $S'' = \{A_3; A_4\}$ are also HSC. In this case, mining S' and S'' are sufficient to capture all the interacting attributes. However, this is a strict condition that need to be relaxed to capture more information as seen in the next definition.

DEFINITION 5.5. Let Γ_k denote the set of all subsets of S with $k-1$ attributes. S is a **Special Combination Of Interest** (or **SCOI**) if it satisfies:-

1. S is a HSC,
2. Exactly one member (say set X) $\in \Gamma_k$ is a HSC and all others are NSC, and
3. $\Delta_{TCI} = TCI(S) - TCI(X)$ is statistically significant at significance level α_{High} .

Let $X = S \setminus \{A_k\}$. Then, it can be easily shown that $\Delta_{TCI} = H(A_k) + H(X) - H(S) = TCI(\bar{X}; A_k)$, where \bar{X} represents a new attribute formed by the joint of all attributes in X . The motivation behind the definition of SCOI is based on the following example. Assume set $S = \{A_1; A_2; A_3; A_4\}$ is a HSC and only its subset $S' = \{A_1; A_2; A_3\}$ is a HSC. If $\Delta_{TCI} = TCI(A_1 A_2 A_3; A_4)$ is significant, A_4 is contributing significantly to the increased correlation information. If we only mine S and not S' , we lose important association information contributed by A_4 *only in combination with S* .

Next, we consider correlations among data attributes (e.g. linkage disequilibrium in genetic data) which can result in redundancy (i.e. presence of overlapping information) among the attribute combinations. Using the property that KWII is negative in presence of redundancy, we have,

DEFINITION 5.6. *Two attributes A_i and A_j are redundant if $Red(A_i; A_j) = \frac{KWII(A_i; A_j; A_j)}{\min\{H(A_i), H(A_j)\}} \leq -\Delta$, where $0 \leq \Delta \leq 1$ is a user specified threshold.*

The definition is based on the fact that if A_i and A_j are redundant, they are in fact interacting, i.e. A_i explains A_j very well. Also A_j completely explains itself (A_j) causing the expression $KWII(A_i; A_j; A_j)$ to have redundant information. The denominator is used to normalize the KWII and is based on the easy to prove fact that $KWII(A_i; A_j; A_k) \leq \min\{H(A_i), H(A_j), H(A_k)\}$.

Compared with the TCI, the KWII is a more valuable information metric because it is a parsimonious measure of association for the attribute combination of interest alone and does not contain contributions from lower-order combinations [16]. However, KWII alone cannot be used to devise an efficient mining algorithm because it takes on both positive and negative values. Only all individual entropies and a joint entropy are needed for one TCI calculation, making it computationally far more tractable than the KWII. The TCI is always non-negative and increases monotonically with increased combination size making it potentially suitable for our mining algorithm. From theorem 4.2, the TCI represents the cumulative synergy present in all subset combinations of the attribute set $\{A_1; A_2; \dots; A_k\}$. Our goal is therefore to use the TCI in our mining algorithm to identify the regions in the combinatorial space (the COI and the SCOI) that contain potentially large correlation information (and therefore high interaction information) and then compute the KWII for the reduced combinatorial space. Thus we shall concomitantly mine attribute sets containing useful correlation information (i.e. TCI) and interaction information (i.e. KWII).

5.2 Mining Strategy Given a maximum order of combinations to explore (K) and a pair of significance levels ($\alpha_{High}, \alpha_{Low}$), our strategy of mining combinations with significant TCI and KWII broadly consists of two steps :-

1. Discover COI and SCOI, and
2. Compute KWII of all K -subsets of the attributes present in combinations mined in step 1.

In step 1, we explore the search space in a breadth-first manner that results in a set enumeration tree as shown in Figure 2. When mining for COI and SCOI, computing the TCI of every attribute set is time consuming, therefore, in the next section we shall develop upper and lower bounds of TCI based on that of its parent/ancestor/sibling nodes in the search space. We further develop pruning strategies using definitions of COI, SCOI and redundancy (definitions 5.4-5.6).

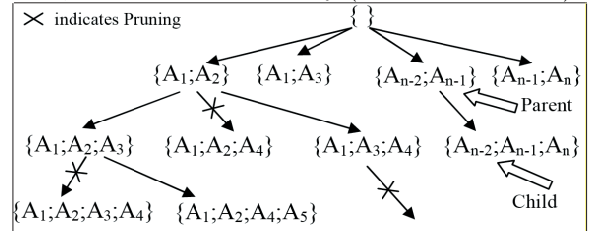


Figure 2: Sample tree enumeration of BFS.

6 Bounds on TCI

In this section, we present results on upper and lower bounds on TCI. The *Value* computation on these bounds shall be used to speed up our mining strategy. In obtaining the upper and lower bounds, we shall assume TCI computations on the attribute set $S = \{A_1; \dots; A_k\} \subseteq \zeta$ unless otherwise stated.

THEOREM 6.1.

$$TCI(S) \geq \sum_{i=1}^k H(A_i) - \frac{1}{2}[H(S \setminus \{A_1\}) + H(S \setminus \{A_2\}) + H(A_1; A_2)]$$

Proof. We can write $H(S)$ is three ways,

$$\begin{aligned} H(S) &= H(A_1|A_2; \dots; A_k) + H(A_2; \dots; A_k) \\ H(S) &= H(A_2|A_1; A_3; \dots; A_k) + H(A_1; A_3; \dots; A_k) \\ H(S) &= H(A_3; \dots; A_k|A_1; A_2) + H(A_1; A_2) \end{aligned}$$

Adding all three, $3H(S) = X + Y$. Here,

$$\begin{aligned} X &= H(A_1|A_2; \dots; A_k) + H(A_2|A_1; A_3; \dots; A_k) + H(A_3; \dots; A_k|A_1; A_2) \\ &\leq H(A_1|A_2; \dots; A_k) + H(A_2|A_3; \dots; A_k) + H(A_3; \dots; A_k) \\ &= H(A_1; \dots; A_k) \text{ (By Chain Rule of Entropy)} \end{aligned}$$

And,

$$Y = H(A_2; \dots; A_k) + H(A_1; A_3; \dots; A_k) + H(A_1; A_2) \\ = [H(S \setminus \{A_1\}) + H(S \setminus \{A_2\}) + H(A_1; A_2)]$$

Therefore, $3H(A_1; \dots; A_k) \leq H(A_1; \dots; A_k) + Y$

$$(6.3) \quad \text{or, } H(S) \leq \frac{1}{2}Y$$

$$\text{So, } TCI(S) = \sum_{i=1}^k H(A_i) - H(S) \geq \sum_{i=1}^k H(A_i) - \frac{1}{2}Y \blacksquare$$

The above theorem computes a lower bound on $TCI(S)$ using entropy from the ancestor nodes. We first use it recursively in computing the upper bound of $H(S)$ (equation 6.3) in a greedy fashion - first obtain its two-attribute subset (say $(A_i; A_j)$) with maximum pair-wise entropy and then recursively compute upper bounds of the entropies $H(S \setminus \{A_i\})$ and $H(S \setminus \{A_j\})$. The upper bound on $H(S)$ is then used to compute the lower bound of $TCI(S)$.

THEOREM 6.2.

$$TCI(S) \leq TCI(S \setminus \{A_t\}) + \min\{H(S \setminus \{A_t\}), H(A_t)\}$$

Proof. We have, $TCI(S) - TCI(S \setminus \{A_t\}) = H(A_t) + H(S \setminus \{A_t\}) - H(S) = H(A_t) - H(A_t | S \setminus \{A_t\}) \leq H(A_t)$. Similarly, $TCI(S) - TCI(S \setminus \{A_t\}) = H(A_t) + H(S \setminus \{A_t\}) - H(S) = H(S \setminus \{A_t\}) - H(S \setminus \{A_t\} | A_t) \leq H(S \setminus \{A_t\})$. Thus, $TCI(S) - TCI(S \setminus \{A_t\}) \leq \min\{H(A_t), H(S \setminus \{A_t\})\}$ ■

The theorem computes an upper bound on $TCI(S)$ using TCI and entropy of its parent node $\{S \setminus \{A_t\}\}$ and $H(A_t)$. The next two theorems are used to compute the upper and lower bounds of the node $\{S; A_j\}$ using entropy of its sibling $\{S; A_i\}$, entropies of individual attributes and conditional entropies. Note that each conditional entropy of form $H(A_i | A_j)$ is given by $H(A_i; A_j) - H(A_j)$.

THEOREM 6.3.

$$TCI(S; A_j) \geq \sum_{t=1}^k H(A_t) + H(A_j) - H(S; A_i) \\ - \min_{t=1}^k \{H(A_j | A_t)\}$$

Proof. By the chain rule of entropy, $H(S; A_j) = H(S) + H(A_j | S)$. But $H(S) \leq H(S; A_i)$ and $H(A_j | S) \leq \min_{t=1}^k \{H(A_j | A_t)\}$ so that $H(S; A_j) \leq H(S; A_i) + \min_{t=1}^k \{H(A_j | A_t)\}$. The result follows by inserting the last inequality in the TCI definition 3.3. ■

THEOREM 6.4.

$$TCI(S; A_j) \leq \sum_{t=1}^k H(A_t) + H(A_j) - H(S; A_i) + \Lambda$$

where, $\Lambda = \min\{H(A_i | A_j), \min_{t=1}^k \{H(A_j | A_t)\}\}$

Proof. By the chain rule of entropy, $H(S; A_i; A_j) = H(S; A_j) + H(A_i | S; A_j)$. But $H(A_i | S; A_j) \leq H(A_i | A_j)$ and also $H(A_i | S; A_j) \leq \min_{t=1}^k \{H(A_i | A_t)\}$, so that $H(A_i | S; A_j) \leq \min\{H(A_i | A_j), \min_{t=1}^k \{H(A_i | A_t)\}\}$, i.e., $H(A_i | S; A_j) \leq \Lambda$. Therefore, $H(S; A_i; A_j) \leq H(S; A_j) + \Lambda$. Also $H(S; A_i; A_j) \geq H(S; A_i)$ giving $H(S; A_j) \geq H(S; A_i) - \Lambda$. The result follows by inserting the last inequality in the TCI definition 3.3. ■

7 Probability Distribution of TCI

In this section, we derive the probability distribution of the TCI using a Taylor series based approximation to the TCI. This shall be used to evaluate the significance of the correlation information of an attribute set.

THEOREM 7.1. *The TCI of an attribute set $S = \{A_1; \dots; A_k\}$ can be approximated as,*

$$TCI(S) \approx \frac{1}{2 \ln(2)} \sum_{a_1, \dots, a_k} \frac{(p(a_1 \dots a_k) - p(a_1) \dots p(a_k))^2}{p(a_1) \dots p(a_k)}$$

Proof. Let $p(a_1 \dots a_k) = \psi_1$ and $p(a_1) \dots p(a_k) = \psi_2$. Let $f(\psi_1) = p(a_1 \dots a_k) \log_2 \left(\frac{p(a_1 \dots a_k)}{p(a_1) \dots p(a_k)} \right) = \psi_1 \log_2 \left(\frac{\psi_1}{\psi_2} \right) = \frac{\psi_1}{\ln(2)} \ln \left(\frac{\psi_1}{\psi_2} \right)$. Using Taylor's expansion of $f(\psi_1)$ about $\psi_1 = \psi_2$, we have,

$$f(\psi_1) = f(\psi_2) + f'(\psi_2) \frac{(\psi_1 - \psi_2)}{1!} + f''(\psi_2) \frac{(\psi_1 - \psi_2)^2}{2!} + \dots$$

Here, $f'(\psi_1) = \frac{\ln(\psi_1) + 1 - \ln(\psi_2)}{\ln(2)}$ and $f''(\psi_1) = \frac{1}{\psi_1 \ln(2)}$. so, that $f(\psi_1) = \frac{\psi_1 - \psi_2}{\ln(2)} + \frac{1}{2 \ln(2) \psi_2} (\psi_1 - \psi_2)^2 + \dots$ Ignoring higher order terms in the Taylor's expansion,

$$TCI(S) = \sum_{a_1, \dots, a_k} f(\psi_1) \\ = \sum_{a_1, \dots, a_k} \frac{\psi_1 - \psi_2}{\ln(2)} + \frac{1}{2 \ln(2) \psi_2} (\psi_1 - \psi_2)^2 + \dots \\ (7.4) \approx \sum_{a_1, \dots, a_k} \frac{\psi_1}{\ln(2)} - \sum_{a_1, \dots, a_k} \frac{\psi_2}{\ln(2)} + \sum_{a_1, \dots, a_k} \frac{(\psi_1 - \psi_2)^2}{2 \ln(2) \psi_2}$$

The first two summations in equation 7.4 sum to $1/\ln(2)$ resulting in theorem 7.1. ■

Note that the expression of TCI is related to the multidimensional statistical χ^2 test [28] defined as,

$$(7.5) \quad \chi^2 = \sum_{i_1, \dots, i_k} \frac{(O_{i_1 \dots i_k} - E_{i_1 \dots i_k})^2}{E_{i_1 \dots i_k}}$$

where the summation is over the cells of the k -dimensional contingency table, $O_{i_1 \dots i_k}$ denotes the observed cell count and $E_{i_1 \dots i_k}$ denotes the expected cell count for cell $i_1 \dots i_k$. The degrees of freedom employed in the omnibus analysis of a k -dimensional contingency table are $df = \prod_{i=1}^k R_i - \sum_{i=1}^k R_i + k - 1$ [28]. Here R_i denotes the count of distinct values that attribute A_i can take. Equating the observed and expected cell counts to the relative frequencies and the cell probabilities, it can be easily observed that,

$$(7.6) \quad \chi^2 = 2 N \ln(2) \widehat{TCI}(A_1; \dots; A_k)$$

where N denotes the total number of samples in the data (i.e. sum of cell counts in all cells of the k -dimensional contingency table). \widehat{TCI} represents the approximation to the TCI derived in equation 6.3. Using theorem 7.1 and equation 7.6 we claim that,

THEOREM 7.2. *The distribution of $\widehat{TCI}(A_1; \dots; A_k)$ can be approximated by a gamma distribution with scale parameter $= 1/(N \ln(2))$ and shape parameter $= df/2$.*

Proof. Let Z be the random variable representing $\widehat{TCI}(A_1; \dots; A_k)$. Using equation 7.6, we can write $\chi^2 = cZ$ where $c = 2 N \ln(2)$ (a constant). Using properties of moment generating functions, the moment generating function of Z is given by

$$(7.7) \quad \begin{aligned} M_Z(t) &= E(e^{tZ}) = E(e^{t \frac{\chi^2}{c}}) = M_{\chi^2}\left(\frac{t}{c}\right) \\ &= \left(1 - \frac{2t}{c}\right)^{-\frac{df}{2}} = \left(1 - \frac{t}{N \ln(2)}\right)^{-\frac{df}{2}} \end{aligned}$$

which is the moment generating function of the gamma distribution with scale $1/(N \ln(2))$ and shape $df/2$. ■

Using theorem 7.2, the $Pvalue$ of an observed TCI value t is given by $Prob(TCI > t)$.

8 Algorithm

Our mining algorithm consists of two stages (as mentioned briefly in section)-(1)Correlation Information Miner (or CIM) followed by (2) Interaction Information Miner (or IIM). The CIM explores the combinatorial space of attribute sets using a breadth-first search (BFS) enumerating a BFS tree where each node represents an attribute set $\{A_i; A_j; \dots; A_k\}$ ($i \leq j \leq \dots \leq k$). Next we describe pruning strategies using the concept of redundancy and bounds on TCI introduced before.

8.1 Redundancy based pruning This pruning strategy is applied to the given data set D before starting the BFS strategy using the redundancy definition 5.6. The goal is to remove redundant attributes thereby reducing the size of the combinatorial space of attribute

associations. It consists of (I) For each attribute $A_i \in \zeta$, compute $Red(A_i; A_j)$ with every other attribute $A_j \in \zeta$. If $Red(A_i; A_j) \leq -\Delta$, store A_j in a list associated with A_i . This step will create a list of attributes redundant with each A_i denoted as $Cover(A_i)$ (which includes A_i). An attribute $A_j \in Cover(A_i)$ is said to be covered by A_i . E.g. if A_1 is redundant with A_2, A_5 and A_8 , $Cover(A_1) = \{A_1; A_2; A_5; A_8\}$. (II) Create a smaller data set D' by greedily selecting attribute A_i with highest cardinality $|Cover(A_i)|$ (i.e, covering the maximum number of other attributes) until all attributes $\in \zeta$ are covered.

8.2 Sample Size based pruning Given attribute set $S = \{A_1; \dots; A_k\}$ and sample size N , $TCI(S)$ and $KWII(S)$ are based on empirically estimated probabilities distributions of the attributes and their combinations from set S . Let the cardinality of the set of attribute values of S be V . The calculated TCI and $KWII$ are often poor estimates when $N/V < 5$ [3]. Therefore, we prune node S when $N/V < 5$ to reduce the chances of discovering false positive associations. For example, to evaluate TCI of $\{A_1; A_2; A_3\}$ where attribute takes 3-values, there should be least $3^3 \times 5 = 135$ instances.

8.3 TCI Bound based pruning For each node S in the search space, we calculate its upper and lower bounds before actual $TCI(S)$. Let $L(S)$ be the maximum of the lower bounds, $U(S)$ be the minimum of the upper bounds and $T(S)$ be the true TCI for node S . Let $P(v)$ be the $Pvalue$ for any value v . Note that as $L(S) \leq T(S) \leq U(S)$, we have $P(L(S)) \geq P(T(S)) \geq P(U(S))$. Assume that we have determined if S is a HSC/MS/NSC. We shall employ the procedures **Handle HSC** and **Handle MS/NSC** described below to handle each case. In the following, in each iteration, $NextLevel$ is a queue that collects nodes to be explored in the next iteration of BFS and Θ is a set of COI/SCOI output by CIM.

(1) **Handle HSC** : Assume that the parent of the given node S is not a COI/SCOI. Using property 2 in definition 5.4, if S is a COI, store S in Θ and add node S to $NextLevel$; otherwise, prune subtree rooted at S as at least one subset of S has redundant correlation information. If the parent of S is a COI/SCOI, using property 2 and 3 in definition 5.5, if S is a SCOI, store S in Θ and add node S to $NextLevel$; otherwise, prune subtree rooted at S as the new attribute present in S (and not in its parent) does not significantly increase the correlation information.

(2) **Handle MS/NSC** : If node S is a MS, S and any superset of it cannot be a COI/SCOI. So simply prune the subtree rooted at S . If it is a NSC, add S to $NextLevel$ to continue the search process.

Based on the TCI bounds, we have the following cases:-

1. $P(U(S)) \leq P(L(S)) < \alpha_{High}$: S is a *HSC*. Use **Handle HSC** to handle it.
2. $P(U(S)) < \alpha_{High} \leq P(L(S)) < \alpha_{Low}$:
3. $P(U(S)) < \alpha_{High}, \alpha_{Low} \leq P(L(S))$:
4. $\alpha_{High} \leq P(U(S)) < \alpha_{Low} \leq P(L(S))$: Compute the TCI $T(S)$. If $Pvalue(T(S)) < \alpha_{High}$, S is a *HSC*, use **Handle HSC** to handle it. Otherwise use **Handle MSC/NSC**.
5. $\alpha_{Low} \leq P(U(S)) \leq P(L(S))$: S is a *NSC*, use **Handle MSC/NSC**.
6. $\alpha_{High} \leq P(U(S)) \leq P(L(S)) < \alpha_{Low}$: S is a *MSC*, use **Handle MSC/NSC**.

Note that actual TCI computations are required only in cases 2,3 and 4 thereby improving computational efficiency. Next we describe the CIM algorithm.

8.4 The CIM Algorithm We assume that CIM uses the data obtained after redundancy removal (section 8.1) for all computations of correlation information and the upper and lower bounds. The inputs are the significance levels α_H for α_{High} and α_L for α_{Low} . Lines 2-8 computes the TCI for every pair of attributes and stores it in *NextLevel* only if the node is a *HSC* or a *NSC*. The *HSC* are collected in Θ to be output. Lines 9-33 explores the combinatorial search space in a breadth-first fashion wherein each node is evaluated to be a *HSC/MS/NSC* and either the subtree rooted at the node is pruned or the search process is continued depending upon the TCI bound based conditions 1-6 outlined above. The sample size based pruning takes place in line 14.

Algorithm CIM($\zeta, \alpha_H, \alpha_L$)

Input: Set of attributes, $\alpha_{High}, \alpha_{Low}$

Output: Θ (set of COI and SCOI)

1. $NextLevel \leftarrow \phi; \Theta \leftarrow \phi;$
2. **for** attribute pair $S = \{A_i, A_j\}$ **do**
3. **if** ($P(TCI(S)) < \alpha_H$)
4. Add S to $NextLevel, \Theta;$
5. **elseif** ($P(TCI(S)) \geq \alpha_L$)
6. Add S to $NextLevel;$
7. **endif**
8. **endfor**
9. **while** $NextLevel \neq \text{empty}$ **do**
10. $CurrLevel \leftarrow NextLevel;$
11. $NextLevel \leftarrow \phi;$
12. **for** each $P \in CurrLevel$ **do**
13. **for** each child S of P **do**
14. **if** not enough samples, goto line 31;
15. Calculate $U(S), L(S), P(U(S)), P(L(S));$
16. **if**($P(L(S)) < \alpha_H$) **do**
17. Handle *HSC* to update $NextLevel, \Theta;$

18. **elseif**($P(U(S)) < \alpha_H \leq P(L(S)) < \alpha_L$) or
19. ($P(U(S)) < \alpha_H, \alpha_L \leq P(L(S))$) or
20. ($\alpha_H \leq P(U(S)) < \alpha_L \leq P(L(S))$))
21. $T \leftarrow TCI(S);$
22. **if**($Pvalue(T) < \alpha_H$)
23. Handle *HSC* to update $NextLevel, \Theta;$
24. **else**
25. Handle *MSC/NSC* to update $NextLevel;$
26. **endif**
27. **elseif**($(\alpha_H \leq P(U(S)) \leq P(L(S)))$ or
28. ($\alpha_H \leq P(U(S)) \leq P(L(S)) < \alpha_L$))
29. Handle *MSC/NSC* to update $NextLevel;$
30. **endif**
31. **endfor** //for each child
32. **endfor** //for each P
33. **endwhile**
34. **return** $\Theta;$

Next we describe the IIM algorithm that is used to compute KWII from the attribute sets output by *CIM*.

8.5 IIM Algorithm Let $\nu \subseteq \zeta$ be the set of attributes present in Θ (combinations output by *CIM*). Let K be maximum order of combinations to be explored. Assuming the sample size to be N and the cardinality of the set of values of the K^{th} order combination to be V , K is chosen such that $N/V \geq 5$. The following algorithm computes the *KWII* of attribute sets of order $\leq K$.

Algorithm IIM(ν, K)

Input: ν (set of attributes present in Θ), K (order of the largest attribute set $\in \Theta$)

Output: Λ (set of combinations and their KWII)

1. $\Lambda \leftarrow$ entropies of all subsets τ of ν s.t. $|\tau| \leq K;$
2. **for** $A_i \in \nu$ **do**
3. **for** each subset X of $\nu/\{A_i\}$, s.t. $|X| < K$ **do**
4. $\Lambda(X \cup \{A_i\}) \leftarrow \Lambda(X \cup \{A_i\}) - \Lambda(X)$
5. **endfor**
6. **endfor**
7. **return** $\Lambda;$

In *IIM*, the array Λ is indexed by attribute combinations. We initialize Λ with entropies of all subsets of ν containing upto K attributes (line 1). For example, with 3 attributes A_1, A_2, A_3 and $K = 2$, $\Lambda(\{A_1\}) = H(A_1)$, $\Lambda(\{A_2\}) = H(A_2)$, $\Lambda(\{A_3\}) = H(A_3)$, $\Lambda(\{A_1, A_2\}) = H(A_1, A_2)$, $\Lambda(\{A_1, A_3\}) = H(A_1, A_3)$ and $\Lambda(\{A_2, A_3\}) = H(A_2, A_3)$. In the end, Λ shall contain negative of KWII values for each attribute combination.

9 Experimental Results

In all our experiments, unless otherwise stated, we have set parameters α_{High} , α_{Low} and Δ to 10^{-8} , 10^{-2} and 0.75 respectively. NIFS [32] was run with parameter

values $\alpha = 0.2$ and $\beta = 0.8$ as used in the paper. In our method, one can set the α 's depending on the experiment and data size, e.g. one would set them conservatively to adjust for multiple comparisons. The Δ can be set to a value > 0.7 depending on how much redundancy one wants to remove from the data.

9.1 Experiment 1 Here, we evaluate the effectiveness of our mining method in detecting attribute associations using a synthetic data set. The data consists of 15 binary attributes and 200 samples. The associations embedded in the data are (1) $A_1 = A_2 \oplus A_3$, (2) $A_6 = A_7 \oplus A_8 \oplus A_9$, and (3) $A_{11} = A_{12} \oplus A_{13} \oplus A_{14}$ where \oplus denotes exclusive-or operation. In addition, noise is added by flipping each of A_1, A_6 and A_{11} with error probability p . We repeat the experiment 100 times. Figure 3A and B show the TCI and top 10 KWII mined by CIM and IIM, respectively for $p = 0.1$. The results are presented graphically as a *spectra* of TCI/KWII values plotted against attribute combinations. Utilizing statistical significance based mining, CIM successfully identifies the embedded associations exactly. The KWII spectra contains the three strongest associations and also other weak associations which can be visually identified as spurious combinations. Figure 3C shows that % of combinations with significant correlation information detected by CIM and is compared with the two methods NIFS[32] and mRMR[25]. The error probability p is varied as 0, 0.1, 0.15, 0.2, 0.25 and 0.3. As NIFS uses hard thresholds, it fails to detect the attribute associations when the strength of an associations varies due to noise in the data. CIM solves this problem by mining with statistical significance levels instead of threshold values. The other method mRMR finds subset of attributes with minimal redundancy among the attributes and a class label attribute. As mRMR requires a class attribute, we have run mRMR separately with: (1) A_1 , (2) A_6 , and (3) A_{11} as the class attribute. However, mRMR performs poorly (even at $p=0.0$) because mRMR uses only mutual information between each attribute and the class to identify the associations.

9.2 Experiment 2 This experiment is modeled after pure epistasis [10] between two SNPs affecting a disease trait. A SNP is a DNA sequence variation in a base pair position at which different sequence alternatives (alleles) exist among individuals in some population. The set of SNPs on a single chromosome of a pair of homologous chromosomes is referred to as a haplotype, and two haplotypes taken together constitutes a genotype. Each SNP usually has two alleles (e.g. A, a) resulting in three genotype values (AA, Aa, aa). In a case-control experiment, the disease trait is binary (0=healthy, 1=diseased). The simulated data in this

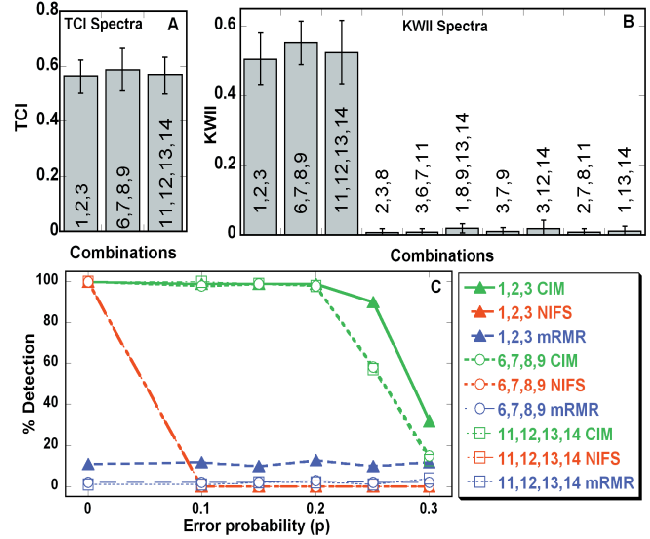


Figure 3: (A) TCI spectra (B) KWII spectra (C) Comparison of CIM with NIFS and mRMR.

experiment consists of 16 discrete attributes: $A_1 - A_{15}$ represent SNPs each having 3 genotypic states and A_{16} representing the disease trait is binary. The data consists of 100 samples of $A_{16}=0$ and $A_{16}=1$ each.

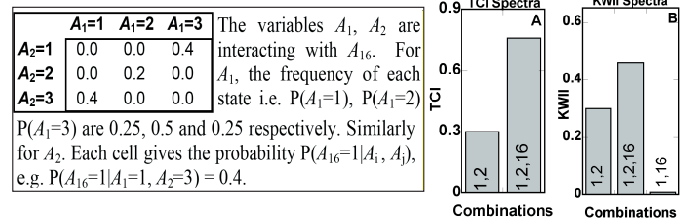


Figure 4: Association model & spectra in Experiment 2.

Associations are created in the data between A_1, A_2 and A_{16} using the model in Figure 4. This results in the following significant combinations:- $(C_1) \{A_1; A_2\}$ and $(C_2) \{A_1; A_2; A_{16}\}$. Note that C_1 is a COI and C_2 is a SCOI. Both CIM and IIM are able to identify both the associations (spectra shown in Figure 4A and B). NIFS identifies only C_1 because it assumes that any superset of a set with strong correlation information contains redundant information. However, in this case, A_{16} can only be identified in combination with C_1 , so that C_2 contains information about A_{16} not present in C_1 . Finally, we run mRMR with A_{16} as the class attribute. Combinations $\{A_1; A_{16}\}$ and $\{A_2; A_{16}\}$ have extremely weak mutual information of 0.008 and 0.003 respectively. As mRMR depends on mutual information between each attribute and the class, it fails to identify any combination involving A_1 and A_2 .

9.3 Experiment 3 In this experiment, the data consists of 16 discrete attributes $A_1 - A_{16}$. A set of complex attribute associations is created involving $A_1, A_2,$

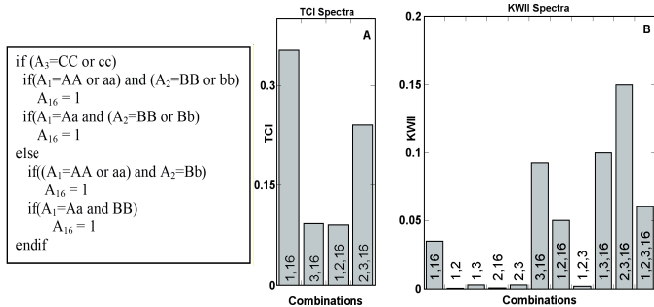


Figure 5: Association model & spectra in Experiment 3.

A_3 and A_{16} . Each of A_1, A_2, A_3 represent SNPs having genotypic states $AA/Aa/aa, BB/Bb/bb$ and $CC/Cc/cc$ respectively. A_{16} stands for a binary disease trait. In addition, redundancy is added by replicating A_1 to A_6, A_2 to A_7 and A_3 to A_8 with 5% error. The data consists of 800 samples of $A_{16}=0$ and $A_{16}=1$ each. The rule that causes A_{16} to be 1 and the TCI and KWII spectra obtained by CIM and IIM are shown in Figure 5. Note that we have effectively removed the redundant attributes (A_6, A_7, A_8) and identified all the interacting attributes. Also observe that the KWII spectra complements the TCI spectra by discovering associations like $\{A_1; A_3; A_{16}\}$ and $\{A_1; A_2; A_3; A_{16}\}$ that are not present in the TCI spectra. Confounded by redundancy, NIFS generates 150 combinations containing attributes $A_1 - A_{10}$ and A_{16} , but does not contain any combination from the TCI spectra identified by CIM as their magnitudes are less than 0.8. mRMR is run with A_{16} as the class attribute and it identifies attributes A_1, A_3, A_{16} in associations $\{A_1; A_{16}\}$ and $\{A_3; A_{16}\}$ but not A_2 because the mutual information $\{A_2; A_{16}\}$ is only 0.0008. These show that CIM and IIM effectively removes redundancy and are capable of identifying a diverse range of attribute associations.

9.4 Runtime Evaluation We have used the following two data sets to evaluate the efficiency of our pruning methods:-(1) Crohns disease dataset [11] is derived from the 616 kilobase region of human Chromosome 5q31 that may contain a genetic variant responsible for Crohns disease by genotyping 103 SNPs and contains 144 case and 243 control individuals. (2) Tick-borne encephalitis dataset [6] consists of 58 SNPs genotyped from DNA of 26 patients with severe tick-borne encephalitis virus-induced disease and 65 patients with mild disease. Figure 6 shows the runtime of our mining method (CIM followed by IIM) under the redundancy based and TCI based pruning strategies as well as when both are applied together and none is applied. For both data sets, the missing values were imputed with the most frequent value for that particular SNP. Sample size based pruning is assumed to be active in all the cases. The number of attributes is varied as follows:

for each data set, from the set of N attributes, a set of K attributes ($K = 10, 20, 30, 40$) is randomly selected and removed from the original data. The experiment is repeated 10 times for each data set and the average runtime for each set of $N - K$ attributes is shown. We observe that, the runtime is least when both pruning strategies are active (green, circles). TCI based pruning (blue, squares) achieves better efficiency than redundancy based (red, rhombuses) pruning in both data sets and the runtime increases exponentially when no pruning is applied. These demonstrates the effectiveness of our pruning methods, as the potential search space is exponential in the number of attributes.

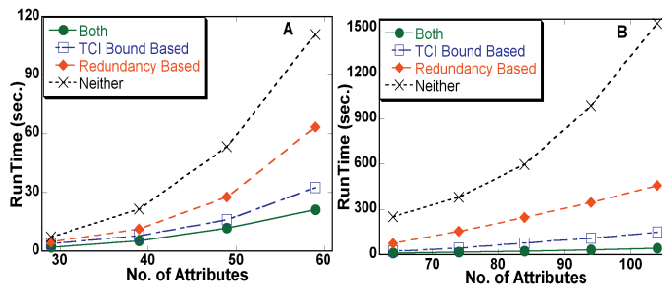


Figure 6: Runtime Evaluation with (A) Tick (B) Crohn's Disease.

9.5 Application to Analysis of SNP-Disease Associations in Chromosome 5 We assess the potential of our mining method (CIM followed by IIM) for identifying key SNPs involved in the causation of Crohn's disease using data set from Daly et al [11]. The 103 SNPs in the data are numbered 0 to 102. Please refer to the 'RunTime Evaluation' section for a detailed data description. Rioux et al. [26] found 11 SNPs ($IGR2055a.1, IGR2060a.1, IGR2063b.1, IGR2078a.1, IGR2096a.1, IGR2198a.1, IGR2230a.1, IGR2277a.1, IGR2081a.1, IGR3096a.1$ and $IGR3236a.1$) with alleles that were associated with risk of Crohn disease. Nine of 11 significant SNPs are present in the data set we analyzed; SNPs $IGR2078a.1$ and $IGR2277a.1$ are missing. For association information mining, subjects and SNPs with missing genotypes are eliminated resulting in 40 SNPs with 58 cases and 92 controls. We perform the following two analyses with the data - (1) Mine the association information in the data without the disease phenotype. (2) Mine the association information such that each combination of SNPs always has the disease phenotype. In our first analysis, we identify three SNPs $IGR2055a.1, IGR2230a.1$ and $IGR3236a.1$ among the combinations with $KWII > 90^{th}$ percentile of the $KWII$ of all combinations obtained. In the second analysis where we take the case/control status into account, the five SNPs $IGR2198a.1, IGR2055a.1, IGR3236a.1, IGR2081a.1$ and $IGR2230a.1$ are found among the

$\{SNP, Phenotype\}$ and $\{SNP, SNP, Phenotype\}$ combinations with $KWII > 90^{th}$ percentile of the $KWII$ of all combinations mined. The TCI and $KWII$ spectra for analysis 2 are shown in Figure 7.

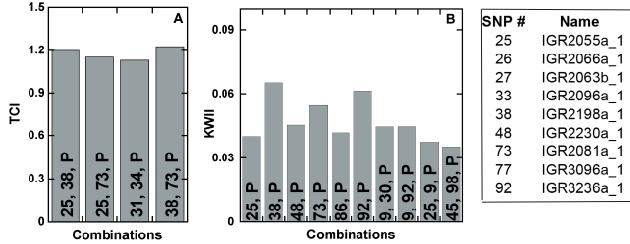


Figure 7: Combinations with TCI and $KWII$ greater than 90^{th} percentile of all the combinations for analysis 2 on Daly’s data. Also disease related SNP numbers and names are given.

On closer examination of the data, we found that due to high linkage disequilibrium in the genomic region examined, SNPs *IGR2066a_1*, *IGR2063b_1* and *IGR2096a_1* belonged to *Cover(IGR2055a_1)* while *IGR3096a_1* belonged to *Cover(IGR2230a_1)* and were pruned during the redundancy based pruning phase of our mining method. However, each of these SNPs is covered by a representative SNP included in the data, as a result, these SNPs and their associated interactions can be easily recovered using the *Cover* data structure after *IIM* completes. For example, consider SNPs 25 (*IGR2055a_1*) and 26 (*IGR2066a_1*). We have $SNP\ 26 \in Cover(SNP\ 25)$, so that for SNP 26, we can get the combinations with high correlation information as $\{26, 38, P\}$ and $\{26, 73, P\}$ and the combinations with high interaction information as $\{26, P\}$ and $\{26, 9, P\}$.

9.6 Application in Feature Selection Here we give an example showing that detecting association information can help with feature selection and classification. Feature selection and classification is a vastly researched and mature area and we do not intend to propose a new feature selection or classification method, instead we show that when we consider multivariate associations among the features, we can improve the accuracy of some existing classifiers. We apply CIM followed by IIM to the following data sets:- (1) Crohn’s disease data from Daly et. al [11] (2) Tick-borne encephalitis dataset [6] and (3) Voting data set [1]. The first two are described in the previous paragraphs. This third data set consists of votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. It has got 16 attributes and 435 samples. Owing to several missing values in the data, the instances containing missing values are eliminated from our analysis resulting in 233 instances. We mine association information such that the class attribute is present in

each combination examined. The table 1 shows the classification results using some traditional classifiers (C4.5, Alternating Decision Tree or ADTree, Random Forest, Naive Bayesian and Classification & Regression Tree or CART). The attributes present in the top ten combinations with the highest $KWII$ are used for Crohn’s disease and Tick-borne encephalitis data sets while the attributes present in the top five combinations with the highest $KWII$ are used for the Voting data set.

	Crohn’s		Tick		Voting	
	All	Sel	All	Sel	All	Sel
J48	39.6	34.9	38.5	31.9	3.9	3.0
ADTree	46.3	35.6	35.2	28.6	4.7	3.1
RF	43.6	38.3	46.2	33.6	4.3	4.3
NB	39.6	34.9	39.6	31.9	8.6	5.6
CART	39.6	35.6	31.9	27.7	3.0	3.0

Table 1: Error rates using various classifiers. ‘Sel’ indicates selected attributes. RF and NB stand for Random Forest and Naive Bayesian respectively

All classifications were performed using Weka [31] with 10-fold cross-validations. Under each data set, the error rates using all the attributes and attributes in selected combinations are shown. The error rates in general decrease with the selected attributes when compared with all the attributes demonstrating that attributes present in combinations with higher association information can improve the classification accuracy.

10 Discussion

In this paper, we have analyzed the problem of mining significant association information between attributes in a data set and have presented novel methods to mine the two types of association information - correlation information and interaction information. Specifically, we have demonstrated and proved the relationships between the two and derived the distributional properties of correlation information that helped us to formulate our mining problem in terms of highly significant, moderately significant and non-significant combinations. We have also derived bounds on correlation information and a redundancy criterion and have developed pruning strategies using them that were found to be extremely effective in pruning the combinatorial search space. Using several experimental and real data sets, we have critically evaluated the effectiveness and efficiency of our mining strategy. For future work, we would like to explore strategies in making our method scalable for handling large number of attributes as commonly observed in genetic data sets.

11 Acknowledgements

We sincerely thank Dr. David Tritchler, Division of Prevention, Ontario Cancer Institute at Toronto,

Canada for helping us with the proof in section 4.

References

- [1] Uci machine learning repository. www.ics.uci.edu/ml/MLRepository.html.
- [2] R. Agarwal and R. Srikant. Fast algorithms for mining association rules in large databases. In *J. B. Bocca, M. Jarke and C. Zaniolo, editors, Proceedings of 20th International Conference on Very Large Data Bases*, pages 487–499, 1994.
- [3] A. Agresti. *Categorical data analysis, 2nd Edition*. Wiley Series in Probability and Statistics, 2002.
- [4] A. J. Bell. The co-information lattice. In *Fourth International Symposium on Independent Component Analysis and Blinds Signal Separation, Nara, Japan, 2003*.
- [5] S. Brin, R. Motwanit, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 265–276, 1997.
- [6] D. Brinza, A. Perelygin, M. Brinton, and A. Zelikovsky. Search for multi-snp disease association. *Proc. Fifth Intl. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS06)*, pages 122–125, 2006.
- [7] N. J. Cerf and C. Adami. Entropic bell inequalities. *Physical Review A*, 55(5):3371–3374, May 1997.
- [8] P. Chanda, L. Sucheston, A. Zhang, D. Brazeau, J. L. Freudenheim, C. B. Ambrosone, and M. Ramanathan. Ambience: A novel approach and efficient algorithm for identifying informative genetic and environmental interactions associated with complex phenotypes. *Genetics*, September 9, 2008.
- [9] P. Chanda, A. Zhang, D. Brazeau, L. Sucheston, J. L. Freudenheim, C. Ambrosone, and M. Ramanathan. Information-theoretic metrics for visualizing gene-environment interactions. *Am J Hum Genet*, 81(5):939–63, 2007.
- [10] R. Culverhouse. The use of the restricted partition method with case-control data. *Hum Hered*, 63:93–100, 2007.
- [11] M. Daly, J. Rioux, S. Schaffner, T. Hudson, and E. Lander. High resolution haplotype structure in the human genome. *Nature Genetics*, 29:229–232, 2001.
- [12] F. Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5:1531–1555, 2004.
- [13] T. S. Han. Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, 46(1):26–45, 1980.
- [14] H. Heikinheimo, E. Hinkkanen, H. Mannila, T. Mielikinen, and J. K. Seppnen. Finding low-entropy sets and trees from binary data. In *Proceedings of KDD*, pages 350–359, 2007.
- [15] K. Ireland and M. Rosen. *A Classical Introduction to Modern Number Theory*. Springer-Verlag, 1990.
- [16] A. Jakulin and I. Bratko. Analyzing attribute dependencies. In *Proc. of Principles of Knowledge Discovery in Data (PKDD)*, Springer-Verlag, 2003.
- [17] A. Jakulin and I. Bratko. Quantifying and visualizing attribute interactions: an approach based on entropy. <http://arxiv.org/abs/cs.AI/0308002v3>, 2004.
- [18] A. Jakulin and I. Bratko. Testing the significance of attribute interactions. In *Proc. of 21st International Conference on Machine Learning (ICML), Banff, Alberta, Canada, 2004*.
- [19] Y. Ke, J. Cheng, and W. Ng. Mining quantitative correlated patterns using an information-theoretic approach. *KDD*, 2006.
- [20] A. J. Knobbe and E. K. Y. Ho. Maximally informative k-itemsets and their efficient discovery. In *Proceedings of KDD*, pages 237–244, 2006.
- [21] Y. Lee, W. Y. Kim, Y. Cai, and J. H. Comine. Efficient mining of correlated patterns. *ICDM*, 2003.
- [22] H. Matsuda. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical Review E*, 62(3):3096–3012, May 2000.
- [23] W. J. McGill. Multivariate information transmission. *IEEE Trans. Inf. Theory*, 4(4):93–111, 1955.
- [24] E. Omiecinski. Alternative interest measures for mining associations. *IEEE Trans. Knowledge and Data Engineering*, 15:57–69, 2003.
- [25] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.
- [26] J. Rioux, M. J. Daly, M. S. Silverberg, K. Lindblad, H. Steinhart, Z. Cohen, T. Delmonte, and et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to crohn disease. *Nature Genetics*, 29:223–228, 2001.
- [27] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
- [28] D. J. Sheskin. *Handbook of Parametric and Non-parametric Statistical Procedures, Second Edition*. Chapman and Hall/CRC.
- [29] M. Tesmer and P. A. Estevez. Amifs: adaptive feature selection by using mutual information. In *Proc. of 2004 IEEE International Joint Conference on Neural Networks*, 2004.
- [30] T. Tsujishita. On triple mutual information. *Advances in Applied Mathematics*, 16:269–274, 1995.
- [31] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, 2005.
- [32] W. W. Xiang Zhang, Feng Pan and A. Nobel. Mining non-redundant high order correlations in binary data. In *VLDB, Auckland, New Zealand, 2008*.
- [33] H. Xiong, S. Shekhar, P.-N. Tan, and V. Kumar. Exploiting a support-based upper bound of pearson’s correlation coefficient for efficiently identifying strongly correlated pairs. *KDD*, 2004.