Discovering Regularities and Irregularities Based on The Distribution of Probability Estimates

Abdul Rahim Halabieh **David K. Y. Chiu**

Computing & Information Science
University of Guelph
Guelph, Ontario, Canada N1G 2W1
abdul@snowhite.cis.uoguelph.ca, dchiu@snowhite.cis.uoguelph.ca

Abstract

This paper proposes a new approach to detect regularities and irregularities in various types of data including multivariate nominal data based on the distribution of the probability estimates of observed events in the data. We first calculate the probability estimates of the events given a sequence of samples. Without assuming the parametric form of the distribution, the distribution of the probability estimates is analyzed. Based on the form of the distribution, relevant events, indicated by significant peaks in the distribution, are detected. Additional statistical evaluations can also be performed. We have evaluated the method using simulated data, a set of user-computer interaction data, and a set of molecular sequences. The results of the analysis show that regularities and irregularities can be discovered flexibly using this method.

Keywords: pattern discovery, regularities-irregularities, multivariate nominal data, pattern recognition in human-computer interaction, molecular sequence analysis.

1 Introduction

Multivariate data analysis has been a main discipline in the study of pattern recognition and image analysis when relationships among data, underlying probabilistic models, and representations of the data are to be known. However a majority of the methods in this area is focused on dealing with multivariate continuous-valued data when there exists a natural ordering of the variable values. One of the reasons is that data of this type can be represented as a point in the Euclidean space, there is a host of techniques including many powerful distance measures can be used. When multivariate nominal data with no natural ordering of the variable values are involved, often analyzing data of this type resort to the analysis of variance as in ANOVA [1, 12], estimating the higher-order probability as in [4, 5, 10], or imposition of an arbitrary scale based on a set of detected weights as in multidimensional or dual scaling [6, 7, 8]. Methods developed from these approaches have been found to be useful in solving problems such as detecting statistical dependency, classification or cluster analysis. Problems arise when the number of variables or the number of possible nominal values of the variables is very large. This is particularly so when these numbers can increase dramatically as in the analysis of ensembles of strings or sequences when positional information is incorporated. Analyzing statistical properties in these cases is usually very difficult and is limited by the sample size in practice.

Without assuming an a priori parametric model of the probability distribution, or even assuming the data distribution to be unimodal, we present here a direct and general approach to analyze multivariate nominal data, especially when the number of possible nominal values is fairly large. This approach can be extended to handle multivariate data of the general type involving mixed continuous and discrete values. We can describe an event here as an n-tuple of nominal values. The observed frequency of a sampling data ensemble can be represented using a high-dimensional contingency table where a large number of cells may be empty.

Our approach is to represent each observed sample by its probability estimate, possibly based on the relative frequency represented in the contingency table. Since the observed data ensemble forms a sample space, we obtain a distribution of the probability estimates. By analyzing the shape of the distribution in terms of relevant "peaks" and "valleys", we identify the corresponding events (or the nominal n-tuple) that show regularities or irregularities with respect to the observations as a whole.

In the next section, we shall present the approach formally. Statistical criteria are proposed to identify "irregularity" - which we identify as observed events that are atypical of the distribution based on their probability estimates as compared to others. These can be "representative" events with significantly high probability estimates, or "unrepresentative" events with significantly low probability estimates. Next, we introduce a discretization scheme to identify "peaks" in the distribution. The events corresponding to the "peaks" represent "regularity" in the observation and are described here as regular or "relevant events" with relatively high probability estimates as compared to those with similar probability estimates. After introducing the approach, a number of examples and experiments are used to evaluate the properties.

2 Description of Method

2.1 Estimating the distribution of the probability estimates

The following description defines the distribution of probability estimates as a function.

Definition: Let $y = \hat{P}(x)$ represent the probability estimate of an observed event x in a sample. Let Y represent the random variable corresponding to y, that is y is the realization of Y. The probability density function f(Y) represents the distribution of the continuous random variable Y. Its graph is the limiting case of the histogram as the sample size increases and the interval decreases in size.

To estimate the distribution of the probability estimates or $\hat{f}(Y)$, we project the probability estimates, $\hat{P}(x)$, into an axis of a new space as in figure 1 (a). The probability estimates of y can be calculated using the relative frequency from the observations. We can estimate f(Y=y) by the number of events in original sequence that have a given probability estimate y and divide it by the number of total observations. f(Y=y) can be estimated as:

$$\hat{f}_1 = f(Y=y) = \frac{k_s}{n}$$

where k_s is the number of samples in the sequence with probability equal to y and n is the total number of observations in the sequence.

To consider only the relative weight of an event among the other possible events, a second method can be used to estimate f(Y=y) by the number of events that have a given probability estimate and dividing it by the number of cells in the contingency table. The estimation can be further adjusted by considering only those cells with probability estimates greater than zero. That is, f(Y=y) can be estimated using a second method as:

$$\hat{f}_2 = f(Y=y) = \frac{k_c}{n}$$

where k_c is the number of cells in the contingency table with probability equal to y and m is the total number of non-zero entries in the contingency table. Note that $\hat{f}_1$ is a sample estimate reflecting the sample distribution and $\hat{f}_2$ is not.

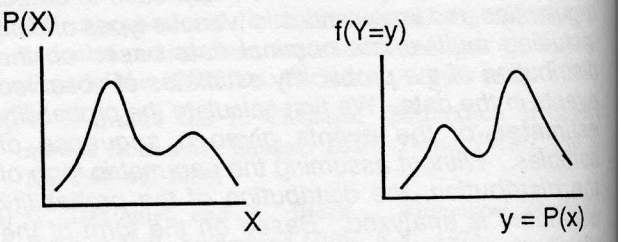

Figure 1 (a): projecting P(X) into the Y-axis of the new space

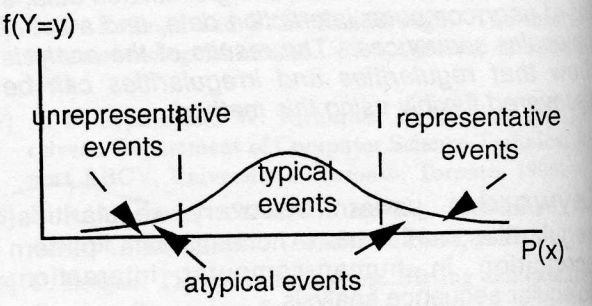

Figure 1 (b): Graphical representation of typical, atypical, representative, and unrepresentative events.

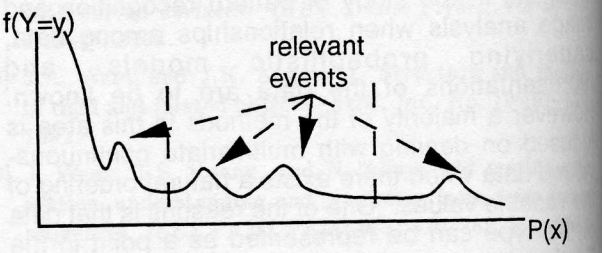

Figure 1 (c): Graphical representation of relevant events.

2.2 Identifying statistically relevant intervals in the probability estimates distribution

2.2.1 Detecting significantly high and low frequency events:

We want to detect events that have occurrences deviate significantly among all the observed events. The events that occur on high and low frequency ends of the distribution of the probability estimates, f(Y), can be considered as "*atypical*" events, that is, from the sample, these events are not "ordinary". Events that occur with medium frequency are then considered as "*typical*" events; see figure 1 (b). Since the Y-axis is continuous, we can calculate the sample mean μ, and the variance δ^2, of f(Y). Using the Tchebycheff's inequality

$$P\left(\left|\hat{f}(Y=y)-\mu\right| \geq k\delta\right) \leq \frac{1}{k^2}$$

for any positive value k, the typical and atypical events can be detected. Note that there is no assumption made on the form of the distribution, f(Y). Furthermore, if unimodal and symmetric distribution is assumed, the inequality can be sharpened as in (Ostle & Mensing) [9]

$$P\left(\left|\hat{f}(Y=y)-\mu\right| \geq k\delta\right) \leq \frac{4}{9}k^2$$

Designing a 1-tail test from the inequality, according to (Chiu '90) [2], we can claim that

$$P\left(\left|\hat{f}(Y=y)-\mu\right| \geq k\delta\right) \leq \frac{4}{18}k^2$$

or approximately

$$P\left(\left|\hat{f}(Y=y)-\mu\right| \geq 2.1\delta\right) \leq 0.05$$

for choosing k = 2.1 and 95% confidence level.

The above inequality implies that there is at least 95% probability that the outcome of f(Y=y) is below $\mu + 2.1\delta$. We call such events with significantly high frequency "*representative*" events since they "dominate" the original sample. Similarly, considering the other side of the tail in the distribution, we can say that 5% events occur on the low end of f(Y). We call such events "*unrepresentative* " events since they occur only a few times (even though they may be of large number) and do not give an indication about the sample; see figure 1 (b).

To summarize: using the proposed statistical test, we can identify from the original sampling ensemble, data that have probability estimates which statistically deviate from the rest of data. Grouping these data into subsets as:

$$S_1 = \left\{e_i : \hat{f}(Y=P(e_i)) \geq \mu + 2.1\delta\right\}$$

and

$$S_1 = \left\{e_i : \hat{f}(Y=P(e_i)) \leq \mu - 2.1\delta\right\}$$

for all events e_i in the sample as sets of "representative" and "unrepresentative" events respectively.

2.2.2 Detecting peaks using Maximum Entropy and Equal Width Partitioning

Interval partitioning is commonly used to estimate the probability density function, such as the use of histogram. Maximum entropy partitioning is a partitioning that minimizes the information loss (or entropy). Another method is to use the equal width partitioning. Maximum entropy partitioning is described as follows.

Let us consider the data represented by Y. The K partition process is defined as one which finds a set of k points on the Y-axis. The set of points will partition the dimension into K intervals, denoted as $\mathbf{R} = \{ R_i : i=1,2,...,K\}$. Let f(Y) be the probability distribution. The process produces a quantization of f(Y) into $P(R_i)$, i=1,2,...,K discrete values to form a discrete probability distribution, denoted by the finite scheme:

$$\Gamma = \begin{pmatrix} R_i \\ P(R_i) \end{pmatrix}, \quad i = 1,2,\ldots,k.$$

The Shannon entropy H associated with Γ is defined as

$$H(R) = -\sum_{i=1}^{k} P(R_i)\log(P(R_i))$$

K-maximum entropy partitioning process is to find the set of boundary points which partition the outcome space into intervals subject to the maximization of the Shannon entropy [3]. The peaks detected then correspond to "*relevant*" events with respect to the whole distribution; see figure 1 (c).

From the maximum entropy partitioning scheme, we can estimate the distribution of f(Y=y) using the maximum likelihood estimation.

3 Evaluation of the method

We have evaluated the method using five examples of data sets. Data set 1 illustrates some of the basic properties of f(Y). Data set 2 illustrates the situation using n independent dice and the situation when using n magnetic (dependent) dice. Data set 3 illustrates the results based on a normally distributed data. Data set 4 illustrates the experimental results using a human-computer interaction data. Data set 5 illustrates the experimental results using a set of molecular sequences.

3.1 Data set 1: Data with different frequencies

The distribution of a data set can be either unimodal or multi-modal. We will look at each distribution separately.

3.1.1 Unimodal data

Consider a data set with one mode (peak) and distribution with (1) small variance, (2) medium variance, and (3) large variance. The distribution of the probability estimates, f(Y), will have different forms. In figure 2 (a), the variance is very small and the probability estimates concentrate near the peak. Correspondingly, it is reflected on the curve f(Y). f(Y) has a relatively low frequency for low y values indicating that the number of low probability estimate events is much smaller than the high probability estimate events. f(Y) gradually increases indicating that the number of high probability estimate events is much greater than the low ones.

For figure 2 (b), the variance is medium such that all probability estimates in the sample occur with roughly the same frequency. The distribution is almost a horizontal line indicating that all probability estimate values have nearly the same frequency in the sample.

For figure 2 (c), the variance of the distribution is very large and the probability estimate values are spread out in the Y-axis. There is a much higher number of low probability estimate values than high probability estimate values as reflected on the curve f(Y). The curve gradually decreases indicating that the number of high probability estimate events is much smaller than the low ones.

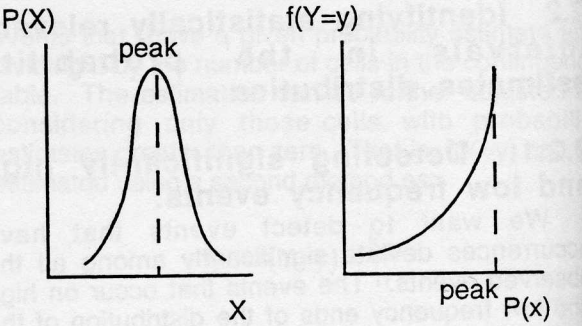

Figure 2 (a): Relationship between P(X) and f(Y) when variance of P(X) is small. Note that f(Y) increases upward.

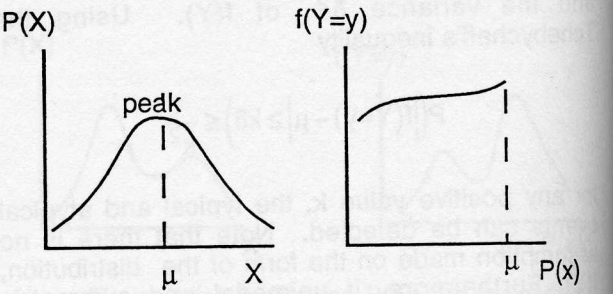

Figure 2 (b): Relationship between P(X) and f(Y) for data "evenly" distributed.

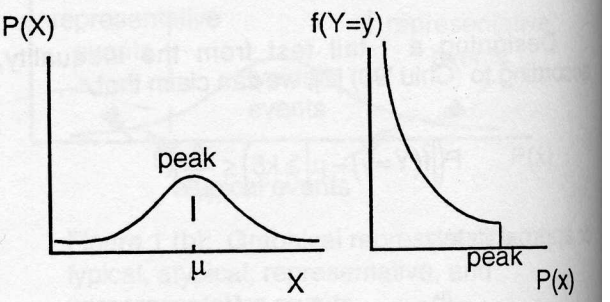

Figure 2 (c): Relationship between P(X) and f(Y) when variance of P(X) is large. Note that f(Y) decreases downward.

3.1.2 Multi-modal data

For a multi-modal data set, the modes will not be reflected easily in the contingency table. In the distribution of the probability estimates, modes with different frequencies will be reflected by different peaks (figure 3). If the two modes have the same probabilities then the peaks in f(Y) will overlap. However, in this case, the peaks can be

differentiated easily from the contingency table.

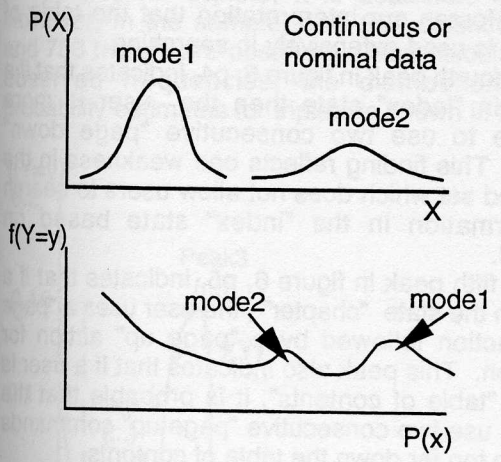

Figure 3: Transformation of multi-modal data to the new space of Y shows the modes of the data.

3.2 Data Set 2: n dice experiment

Suppose that we have n dice. Each die is fair, the outcome of one die does not affect the other, and the dice are said to be independent. The experiment is to throw the n-dice many times. Since the dice are independent, the estimated probability of an event by throwing the dice once is $(\frac{1}{6})^n$. The sample distribution of the probability estimates is unimodal and similar to that in figure 4.

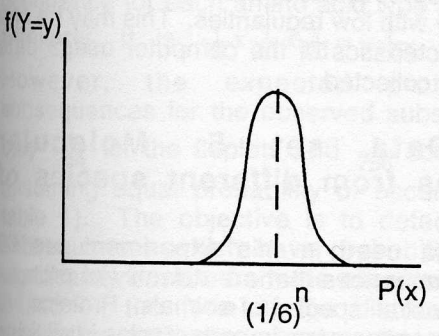

Figure 4: Distribution of the probability estimates when throwing n independent dice in an experiment.

Now consider magnetic dice. That is some of the outcomes in one die will affect the outcomes of other dice, then the distribution of the probability estimates will be multi-modal and similar to that in figure 5. Note that the first peak (on the left) represents outcomes that are not affected by the magnetic force and the second peak (on the right) represents outcomes that are.

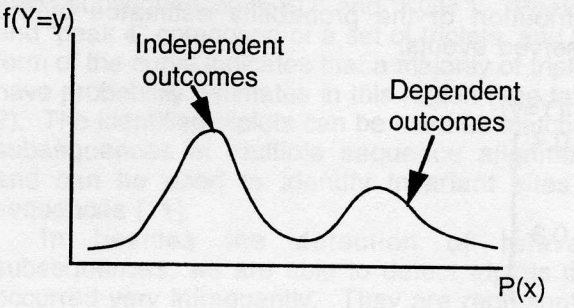

Figure 5: Distribution of the probability estimates when throwing n magnetic (dependent) dice.

3.3 Data set 3: Normally distributed data

This data set is generated with outcomes normally distributed. The variance, denoted as δ^2, is equal to 0.25 and the mean, denoted as μ, is equal to 0. The distribution function of the variable is described as

$$f(x) = \frac{1}{\delta\sqrt{2\pi}} e^{-x^2/2\delta^2}$$

We can estimate f(Y=y) by the area or by counting the number of unit-size squares under the curve using equal width intervals. For this data, the number of low probability estimates is much smaller than the high probability estimates, since the variance is small (i.e. δ^2 equals to 0.25) and the probability estimates concentrate near the peak of the distribution of f(x).

3.4 Data set 4: Computer usage Data in on-line book hypertext environment

The data are computer-traced user operations from usage studies of interactive software. For each command the user issues, the command of the user and the state in the system are recorded. There are 64 different actions such as the "page down" and "page up" actions and 31 different states (e.g. in "table of contents" or "preface"). We analyzed a sequence of 17,499 user commands in terms of the relationship of previous user command, state and current user command. An event is then represented as a 3-tuple (previous action, state, current action).

The 3-dimensional contingency table involves 31x64x64 = 126,976 cells. After generating the

contingency table, the distribution of the probability estimates is shown in figure 6 representing the distribution of the probability estimates of the observed events.

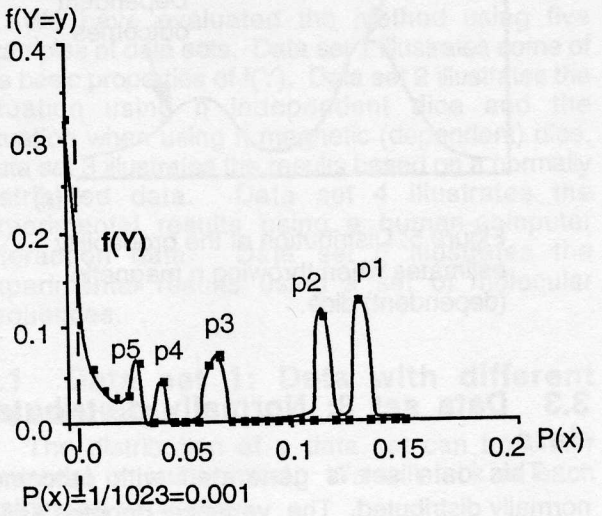

Figure 6: Distribution of the probability estimates of the computer usage data. (data set 4).

From the shape of the distribution shown in figure 6, we identify five peaks representing seven events to be relevant. The distribution shows that most event occurrences are those found from the peaks and that the occurrences of the other events are very infrequent. The usage of commands in different states is mainly concentrated around the detected relevant events.

The event corresponding to the right most peak in figure 6, p1, indicates that if a user is in state "chapter" searching for a piece of information, the user is most likely to use two "page down" commands for traversing. From this finding, the system designer can design a new command that is more appropriate in such a state.

The event corresponding to the second peak from the right in figure 6, p2, indicates the occurrence relationship between two consecutive "cursor down" actions in state "table of contents" as a navigation technique. Since cursoring down to a specified subject is usually done when the target is seen by the user, this peak indicates that the table of contents has been used extensively in searching.

The third peak in figure 6, p3, indicates that if a user is in the "chapter" state then it is highly probable that two "page up" commands are used to look for a piece of information, indicating recovering from going too far down within a "chapter" environment is common. Also, p3 indicates that if a user is in "table of contents" then the user uses two consecutive "page down" commands frequently. This reinforces our interpretation that the table of contents is used extensively in searching.

The fourth peak in figure 6, p4, indicates that if a user is in "index" state then the user is more probable to use two consecutive "page down" actions. This finding reflects one weakness in the command set which does not allow users to search for information in the "index" state based on keyword.

The fifth peak in figure 6, p5, indicates that if a user is in the state "chapter", the user uses a "page down" action followed by a "page up" action for correction. This peak also indicates that if a user is in state "table of contents", it is probable that the user will use two consecutive "page up" commands for going too far down the table of contents.

Considering the two states "chapter" and "table of contents", we find that paging down within those states happened more often than paging up. Moreover, from the graphical representation of the distribution of the probability estimates (figure 6), we see that the commands associated with the state "chapter" is farther to the right in the curve indicating higher usage frequency than in the "table of contents". This finding indicates that the size of the "table of contents" state is much smaller comparing to the "chapter" and therefore, the "page down" action was more frequently used in the state "chapter" than in the state "table of contents". Furthermore, because of the slope of the distribution of the probability estimates generally goes downward, this data set is found to be very noisy with low regularities. This may be one of the characteristics of the computer usage data that we have collected.

3.5 Data set 5: Molecular sequences from different species of animals

The data used in this experiment are 63 myoglobin sequences that are a family of proteins found in almost all species of animals. Proteins are molecules that play many important roles in all living organisms. A protein is made up of a chain of smaller building blocks known as amino acid residues. There are 20 types of amino acids. Different combinations of amino acids sequences define different proteins.

All sequences used in this experiment have 153 amino acid residues. The data was analyzed by studying subsequences of different lengths, in particular length two (duplets) and three (triplets) amino acid units. In total, there are 9,576 observed duplets and 9,513 observed triplets in the sequence ensemble. For duplets, in principle, the

number of possible distinct subsequences is 20x20 = 400 and for triplets, it is 20x20x20 = 8,000. However, in the sample, only 254 distinct duplets and 756 triplets are observed. After calculating the observed frequencies, the distribution of the probability estimates for triplets is shown in figure 7.

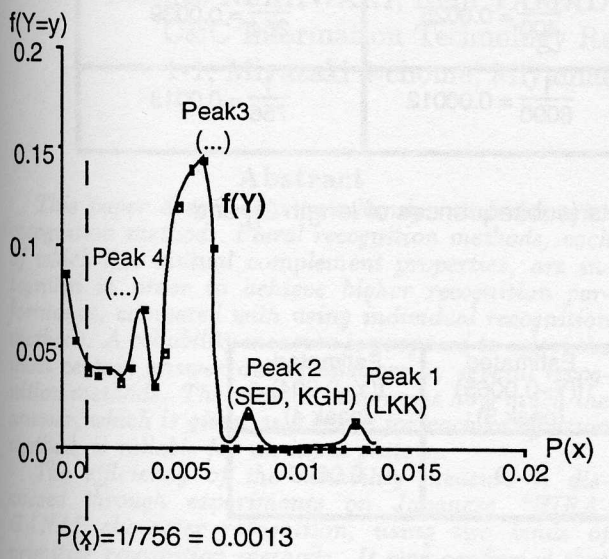

Figure 7: Distribution of the probability estimates of triplets, subsequences of length 3. (data set 5).

For duplets, the expected probability is $\frac{1}{20}x\frac{1}{20}=\frac{1}{400}$ assuming equal probability of occurrence for each amino acid type at a position. Similarly, it is $\frac{1}{20}x\frac{1}{20}x\frac{1}{20}=\frac{1}{8,000}$ for triplets. However, the expected probability of subsequences for the observed subsequences is only $\frac{1}{254}$ for the duplets and $\frac{1}{756}$ for the triplets, assuming equal probability of occurrences (see table 1). The objective is to detect groups of subsequences which have probability estimates significantly deviate from the expected probability. Based on the form of the distribution of the probability estimates, such subsequences are represented by "peaks" on the curve and can be easily identified. By examining figure 7, we see that "peak 1" occurred with probability equals to 0.013, and this probability significantly deviates from the expected value of $\frac{1}{756}$ = P(x) = 0.0013. The triplet represented by this peak 1 is found to be the amino acid subsequence Leucine(L), Lysine(K), and Lysine(K) denoted as "LKK". "Peak 2" in figure 7, with P(x) = 0.0091, also occurred with probability deviates significantly from the expected value of $\frac{1}{756}$. Peak 2 representing two triplets (the amino acids subsequences "SED" and "KGH"). "Peak 3" and "peak 4" composed of a set of triplets, and the form of the curve indicates that a majority of triplets have probability estimates in this region (see table 2). The identified triplets can be used as matching subsequences in multiple sequence alignment, and can be used to identify invariant sites in sequences [11].

In besides the detection of relevant subsequences, we are able to detect triplets that occurred very infrequently. They are represented by triplets that exist to the left of the dashed line in figure 7. The dashed line represents the position of the expected probability of a subsequence assuming equal probability of occurrences for the subsequences observed. These triplets are infrequently observed or unrepresentative of the whole observed data. One possibility is that changes due to insertion, deletion, or modification may occur in these regions.

4 Conclusion

In the proposed method, regularities and irregularities are detected from a sample of observations based on the distribution of the probability estimates. Our method detects relevant events represented as subsequences without any a priori knowledge about the parametric form of the distribution or the domain. Furthermore, our method is able to handle multi-modal data. The process can be achieved in three phases. First, the probability estimates for all the possible events are calculated. Then, the distribution of the probability estimates is plotted and evaluated. Finally, regularities and irregularities are detected represented by peaks in the distribution as well as significantly high and low occurring events. The method has been evaluated using simulated data, a set of computer-usage data, and a set of molecular sequence data with very good results.

Acknowledgments

We would like to thank Dr. Michael Zuker of the National Research Council of Canada for providing us with the myoglobin sequences used in the experiment. Also, we would like to thank professor Tom Carey of the University of Guelph and Dr. Don Edwards of IBM for the fruitful discussion. This work is supported by an operating grant of the National Sciences and Engineering Research Council of Canada, a grant from the University Research Incentive Fund of Ontario, and IBM Canada Ltd.

	Total number of samples	Number of possible distinct subsequences	Number of observed distinct subsequences	Expected probability of a subsequence assuming equal probability of the occurrences for the amino acids	Expected probability of a subsequence assuming equal probability of occurrences for the subsequences observed
Duplets	9576	400	254	$\frac{1}{400} = 0.0025$	$\frac{1}{254} = 0.0039$
Triplets	9513	8000	756	$\frac{1}{8000} = 0.00012$	$\frac{1}{756} = 0.0013$

Table 1: Observed and expected values for duplets (subsequences of length 2), and triplets (subsequences of length 3).

	Estimated f(Y≈0.013) (peak 1)	Estimated f(Y≈0.0091) (peak 2)	Estimated f(Y≈0.0065) (peak 3)	Estimated f(Y≈0.004) (peak 4)
Triplets	0.0126	0.0162	0.49	0.0686

Table 2: Estimated f(Y≈y) or the proportion of samples having $\hat{P}(x)$ significantly deviating from the expected value.

References

[1] Bock, R. D. Multivariate Statistical Methods in Behavioral Sciences. New York: McGraw Hill, 1975.

[2] Chiu, D. K. Y. (1990). Analysis of Discrete-Valued Data Based on Dependence Tree Transformation, University of Guelph, Dept. of Computing & Information science, internal report.

[3] Chiu, D. K. Y., Wong, A. K. C. and Cheung, B. (1991). Information Discovery through Hierarchical Maximum Entropy Discretization, Knowledge Discovery in Databases AAAI press/MIT press 1991. Editors: G. Piatetsky-Shapiro and W. J. Frawley.

[4] Chiu, D. K. Y., Wong, A. K. C. (1986). Synthesizing knowledge: A Cluster Analysis Approach Using Event Covering. IEEE Transactions on Systems and Cybernetics, Vol. **SMC-16**, No. 2, March/April 1986.

[5] Chow, C. K. and Liu, C. N. (1968). Approximating Discrete Probability Distributions with Dependence Trees. IEEE Transactions on Information Theory, **14**: 462-467. May 1968.

[6] Hayashi, C. Multidimensional Quantification - with the applications to analysis of social phenomena. Annals of The Institute of Statistical Mathematics, 1954, **5**, 121-143.

[7] McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models, Monographs on Statistics and Applied Probability 37, Chapman Hall.

[8] Nishisato, S. and Nishisato I. (1984). An Introduction to Dual Scaling. Microstats, Toronto, Canada, 1984.

[9] Ostle, B. and Mensing, R.W. (1975). Statistics in Research (the Iowa State University Press).

[10] Siochi, A. C. and Hix, D. (1991). A study of computer-supported user interface evaluation using maximal repeating pattern analysis. In Proceedings of CHI '91, Conference on Human Factors in Computing Systems. ACM, New York, 1991, pp 301-305.

[11] Waterman, M. S. and Jones, R. (1990). Consensus Methods for DNA and Protein Sequence Alignment. In Enzymol. **183**, 221-237.

[12] Van de Geer, J. P. Introduction To Multivariate Analysis For The Social Sciences. W. H. Freeman and Company, San Francisco, 1971.