Abstract

We survey parts of bioinformatics theory with respect to DNA chip microarray data analysis. First, we outline information structures and bioinformatics itself. Next to it, we describe so called fuzziness and we show generalized logical connectives which are usable for data preprocessing and structuring. Finally, we describe several classes of aggregative operators.

1 Introduction

There are three main structural states of physical matter with respect to its organization: solid, liquid and gas phases. Their classical representants are ideal crystal, ideal fluid and ideal gas respectively. They are depicted at Figure 1. While crystal has fixed and regular structure, gas has random and dynamical structure. Organization of fluids lies between the two extremes. In case of solid state, we deduce all the structural properties of the matter from just one point of it. In case of liquid state, our deduction is limited to a bounded region. We can not deduce anything on distinct parts of matter in case of gas state.

Sometimes, live matter (i.e. organisms) is put in line with fluids. It seems to be rational, since both structures are partially regular. However, there are
some controversies. First, organisms are not just spread fluid matter. Second, there are several patterns for the "middle" setting. We sketch three possible structures at Figure 2. The case A is for partially sublimated matter - if we are lucky, we can deduce investigated properties to large part of the matter. However, in adverse situation, we can not deduce at all. The case B is for fluids and they were mentioned above. The case C is for so called organismal matter. We can deduce just small amount of matter properties from one point knowledge. However, as we investigate more points in the matter, we can deduce much more - not just on bounded surroundings of the investigated points. It is usually the case we assume to be the interesting one. And we believe, it is the case for organisms. Nevertheless, we do not say that it is specific property for living organisms.

![Figure 2: Fluid-like forms of organization](image)

Features being dashed at Figure 2, case C, are covered inside investigated matter. They characterize particular objects, but we do not know the features a priori. The task is to unravel the features. Since the features are too complex and diversified to be covered by a few formulas, we try to spring them by data mining methods. Usually, our work is separated into three parts. First, theoretical algorithms have to be invented. Second, we have to implement the algorithms into software. Third, programs are used on biological data. We focus to the first part in this survey. Especially, we concentrate on use of Hájek's observational calculus and fuzzy logic.

2 Fuzzy logic and bioinformatics

Fuzzy logic [1] is fruitful of structures which can be used for data mining. Unfortunately, the word of "fuzzy" is used for many different ideas [6]. First, we use the notion of fuzzy as is formalized in mathematical fuzzy logic: i.e. logic of comparable truth values. Second, bioinformatical data [1] we focus on, have their values in real intervals. It means that value e.g. 0.5 is for actual half-large variable. For example, one variable can be age: people can range from very young (value $\approx 0.1$), somewhat young (value $\approx 0.3$) to very old (value $\approx 1.0$) ones, see example at Figure 3.
It is not necessary to have linear dependence of a fuzzy value on the real quantity. In case of bioinformatical data, the dependence frequently contains logarithmical transformation. One reason for it is gaining distribution of data which is more symmetrical and normal like.

Contrary to the above case of real continuous data, there are situations with crisp (i.e. two valued - yes/no) data when the meaning of fuzziness is used too. For example, the crisp variable can be a win in a future with its fuzzy value expressing the chance or our hope to win, see at Figure 3. Fuzzy variables which are used for description of such situations, are just measures of probability or believe that investigated crisp data occur. It is notable to say that we do not use fuzziness for such two valued data since one just expresses value of uncertainty there.

We develop methods for biological data that can usually have their values greater or lesser than a middle value. This is motivated by gene expressions. Values of expression are by default viewed as either being in a middle region or altered ones. In case of alteration, the values can be greater (i.e. activated expression) or lesser (i.e. inhibited expression).

Some common examples can be temperature or favor of cup of tea. In case of cup temperature, the tea can have middle temperature - it is neither warm nor cold, it can be cold, it can be warm. Likewise, the tea favor can be as negative (dislikes), neutral or positive (likes). It is shown at Figure 4.

It is natural to use interval of $[-1, 1]$ to express such values. In fact, we use pairs of values for it. It can be gained by usage of generalized logical connectives. The new connectives, say plications, are extension to implication and coimplication as uninorms are extension to t-norms and conorms. It means that in case of a plication, say $P(x, y)$, it generally holds neither $P(x, y) = 1$ for $x \leq y$ nor $P(x, y) = 0$ for $x \geq y$. The new connectives can be used not only for pairs of values on single properties, but they can be reused for general pairs. In such a case, they can express time changes. It is useful tool for time series data and it plays role of time differentials.

Together with it, we reuse principles invented as monadic observational predicate calculus [3, 4]. It has two subsequent parts. Particular measured properties
are used as logical formulas and they are combined by logical connectives. Next to it, generalized forms of quantifiers are evaluated on pairs of formulas to check their connections. It can be viewed as counting on a relational table:

<table>
<thead>
<tr>
<th></th>
<th>var 1</th>
<th>var 2</th>
<th>...</th>
<th>var M</th>
</tr>
</thead>
<tbody>
<tr>
<td>obj 1</td>
<td>0.3</td>
<td>0.8</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>obj 2</td>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>obj N</td>
<td>0.1</td>
<td>0.5</td>
<td>0.4</td>
<td></td>
</tr>
</tbody>
</table>

The exemplary table above shows starting point for observational calculus (on fuzzy data) computing. Separate columns are for particular variables, for example genes or cups of teas. Separate rows are for particular objects, we measure the variables on. They can be patients or drinkers. Filled values (set into interval [0, 1]) can express amount of gene activation / inhibition or tea positive / negative favor, respectively. We look for rules that say e.g. ”who likes tea of kind 1, dislikes tea of kind 2”, ”when both genes 1 and 2 are activated then gene 3 is activated too”.

Combination of variables is done by connectives of fuzzy logic. Since amount of variables in bioinformatics (i.e. genes) is rather big, it is necessary to cluster them during computations. It is not disadvantage. It is known that groups of genes behave similarly and to find the groups is one of tasks of bioinformatics. Evaluations are done by so called generalized quantifiers. They combine ideas of classical quantifiers and ideas of statistical estimators and tests [5]. They can be, for example, estimates of quantiles (of holding a formula) or tests for e.g. 0.9 value of them on a value of significance.

### 3 Feature aggregation

When we have found and enumerated relevant rules we may want them to combine to express a final value which describe investigated system. The value of the object in the interest can be similarity to another (complex) object,
inclination of a relevant gene to be activated or inhibited, or favor of the prepared tea.

We generally have pieces of evidence for both greater final values and lesser final values. Their combination should behave as uninorms. It means that combination of two positive values should tend to be greater, combination of two negative values should be lesser, and combination of one positive and one negative value should lie between them. We can describe such behavior as acting of individual rules on the final value that is glued onto one end of a spring, the second end of the spring is glued to zero value. We call such an operator a dinorm, an example is at Figure 5.

![Figure 5: Dinorm example](image)

We need continuity, rather uniform one, to have stable aggregative operators. However, it is impossible for uninorms. In fact, uninorms have unnatural behavior on combination of two opposite extreme values: it must be an extreme too. It can be overwhelmed by abandoning associativity, either weak or strong. It is not so bad since e.g. (arithmetical) mean is not associative too. We just can not separate the final operator into recursive action of one (associative) binary operator.

Still, we can state less conditions (than recursiveness) on reducibility of the operator. The operator may be, for example, separable into two (several) associative operators. In such a case, we say that the operator obey weak non-associativity. This imitates double values in preprocessing and formula combination steps: first, we combine separately positive and negative values by conorms, and second, we combine the two result values by coimplication (of the lesser one to the greater one). Generally, we do not suffer from lack of associativity since it is not required for aggregation operators - we do not use...
them as logical connectives. We usually want to have evaluated the power of our result from statistical point of view. Since we have an amount of both objects and rules, we can use some multidimensional methods, e.g. bootstrapping. It yields strength and plausibility of localization of the final value on whole [-1, 1] interval. It means that we can state e.g. that the final result value is greater than or equal to 0.5 with a value of significance, and it is greater than or equal to 0.3 with a greater value of significance.

References


