データ取得機構（データクレンジング）　　2018/11/15

データは機械学習によって人工知能に取り込まれ、利用されていきます。そのとき、基本となることは、データが有効な物であることです。雑音で本質的な情報が隠された物であってはなりません。従って、データ取得時にデータを評価して、有効な物のみを機械学習に取り込むようにします。これがデータ取得時のデータクレンジングです。

データは基本的にセンサー群から得られる（key,value）型の値です。そうして、時系列を作ります。時系列の発生という場の状況（文脈）と時系列の値の様子から有効なデータかどうかを判定し、なおかつ、どうゆう状況の下での値なのかと言うことを、把握するような形式で、データを取り込むのが、後々の利用という段階で、有効になる秘訣だと思われます。

音声は１次元の時系列のデータですが、これの解析では音素というものが重要になります。一般的に、音素のような断片的特定パターンのチャンク群で、解析していくことになると思います。ただ、そのチャンク（データ要素）は多くのデータを収集する中で、交差法とクラスタリングで、獲得できる物です。なにはともあれ、値の強さと強さの変化と周波数の強さと周波数の変化が全てのデータ解析の基本で、これを纏めて得られるのが、データ要素なわけです。

ということで、データ要素というチャンクとデータ要素の集まりのチャンクが構成されていきます。それはクラスタリングと交差法で同じパターンが作るチャンクの時系列ネットワークを構成していくことで得られます。オートマトンプールを構成していくのです。そこで、チャンクを得ていく。チャンクは、意味のある単位であり、入れ子構造、交差構造をして存在し、正常値、異常時値、雑音（欠測値、異常値）の弁別が中核のまとまり単位（チャンク）となります。この弁別は教師有り共起学習によるもので、教師はセンサーがおかれた環境の評価によります。自動で教師データは収集できる場合が多いです。自動運転もそんな環境を提供する応用分野です。

繰り返しますと、チャンクは交差法によってオートマトンを作っていきます。データ要素の連続が同じ所を細かなチャンクとするのです。これによって、さらにデータの傾向を把握することができるようになります。

データの意味は、データ発生の時間空間事象という環境情報によって与えられます。その意味とチャンクとの関係も交差法によってより精密に意味づけされることになります。そうして得られたデータは、機械学習に適した綺麗なデータとなっています。

おわり