А.В.ВОЛОШКО, Т.Н.ЛУТЧИН

СТАТИСТИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ ИНФОРМАЦИОННЫХ СИГНАЛОВ НА ОСНОВЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ

Введение

Стремительное развитие информационных технологий привело к необходимости создания таких методов анализа сигналов, которые позволили бы не просто перечислять характерные частоты (масштабы) сигнала, но и получать сведения об определенных локальных координатах проявления этих частот. В настоящее время эффективным методом для анализа и обработки нестационарных (во времени) или неоднородных (в пространстве) информационных сигналов различных типов является вейвлет - анализ.

Вейвлетами называют масштабирующую $\varphi(\mathrm{t})$ и вейвлетную $\psi(t)$ функции, растяжение / сжатие (масштабирование) и сдвиги которых образуют базисы для представления сигналов в виде функционального ряда [1]:

$$x(t) = \sum_{k=-\infty}^{\infty} c_0(k) \cdot \varphi_{j0,k}(t) + \sum_{j=1}^{J} \sum_{k=-\infty}^{\infty} d_j(k) \cdot \psi_{j,k}(t), \quad (1)$$

где $c_0(k)$ и $d_j(k)$ - коэффициенты разложения сигнала по масштабирующей и вейвлетной функциям соответственно;

j - масштаб;

k - сдвиг базисных функций.

Первое слагаемое является аппроксимирующим (грубым) приближением сигнала в масштабе j_0, второе – детализирующим.

Постановка задачи. Метод формирования признаков распознавания, на основе которых будут рассмотрены вопросы статистической кластеризации, предложен в [2,3]. В вейвлет - спектре, сформированном на основе вейвлет - пакетов [4], усредняется мощность рассчитанных вейвлет - коэффициентов в пределах каждой субполосы разложения. Усредненные коэффициенты нормируются и в соответствии с их местом в общей пирамиде вейвлет - пакетов слева направо и сверху вниз преобразуются в вектор признаков распознавания. То есть, в качестве первичных признаков распознавания выступает упорядоченная последовательность средних мощностей вейвлет - коэффициентов по субполосам.

Определение средних мощностей вейвлет - коэффициентов в каждой субполосе производится по следующему выражению [4]:

$$\overline{P}_{m,n} = \frac{\sum_{i=n\cdot\frac{N}{2^m}}^{((n+1)\cdot N/2^m)-1} (k_{m,n}(i))}{\frac{N}{2^m}}, \quad (2)$$

где m - шаг разложения;

N - общее количество аппроксимирующих и детализирующих коэффициентов на m–ом шаге разложения;

n - порядковый номер коэффициента на определенном шаге разложения;

$k_{m,n}$ - вейвлет - коэффициенты.

Вектор признаков, сформированный на базе рассчитанных средних мощностей вейвлет - коэффициентов по субполосам, представляется как $Y = (\overline{P}_{00}, \overline{P}_{10}, \overline{P}_{11}, \overline{P}_{20}, ..., \overline{P}_{m,n})$.

В связи с тем, что вектор признаков описывается однородными коэффициентами, приводящими к равномерности оси пространства, то с практической точки зрения (простоты меры сходства и дальнейших вычислительных процедур) в качестве меры сходства для кластеризации выбирается евклидово расстояние [5] между объектами с минимальным и максимальным средним значением P_{ci}.

В результате чего получены выборки $\rho\,(\overline{P}_{m,n\,i}, \overline{P}_{m,n\,j})_n$:

$$\rho\,(\overline{P}_{m,n\,i}, \overline{P}_{m,n\,j}) = \overline{P}_{m,n\,i} - \overline{P}_{m,n\,j},$$

где n - количество рассматриваемых рядов;

$\overline{P}_{m,n\,i}$ и $\overline{P}_{m,n\,j}$ - объекты сравниваемых рядов.

Полученные значения расстояний образуют заметно различимые группы (рис. 1).

Результаты расчета евклидовых расстояний наглядно демонстрируют величину расхождения между рядами, что позволяет оценить возможность проведения в дальнейшем кластеризации, а также обосновать ее результаты.

Для проверки закона распределения случайных величин, входящих в вектор признаков, рассчитано математическое ожидание (среднее значение ряда) и разброс (среднее отклонение) по известным формулам:

$$P_{ci} = \frac{\sum P_{m,n}}{n},$$

$$\delta_{ci} = \sqrt{\frac{\sum (\bar{P}_{m,n} - P_{ci})^2}{n-1}}, \qquad (3)$$

где n – количество значений ряда;

P_{ci} и δ_{ci} - среднее значение и среднее отклонение соответственно выборки данных.

Результаты вычислений приведены в табл.1.

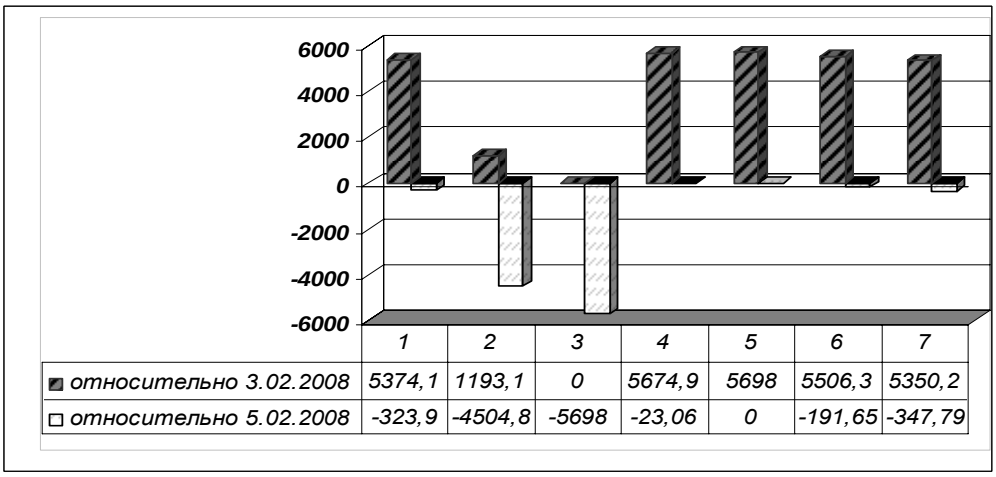

Рис.1. Евклидовые расстояния между объектами

Характеристики, которые целесообразно использовать для поставленных задач, предварительно оцениваются. Так как кластеризация является многомерной статистической процедурой, позволяющей формировать группы из сравнительно однородных исходных данных, то в его качестве таких данных можно принять средние значения выборок для рассматриваемых суток (рис.2).

Таблица 1

Средние значения и средние отклонения для рассматриваемых суток

№ суток	P_{ci}	δ_{ci}
1	2943,848	14031,542
2	2274,796	8013,179
3	1914,709	6168,584
4	2603,249	13912,409
5	3040,636	14552,2
6	2921,272	14172,124
7	2865,889	13847,83

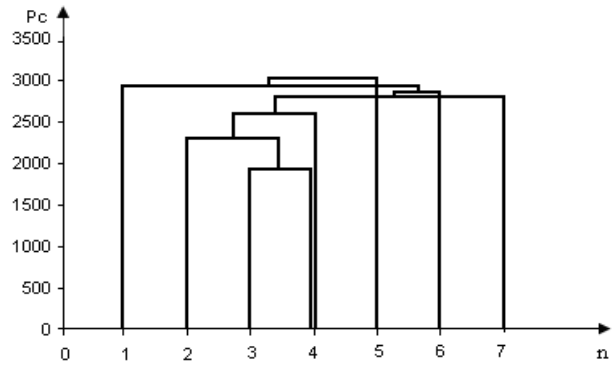

Рис. 2. Дендограмма средних значений выборки данных

Представленная дендограмма основана на средних значениях энергопотребления для рабочей недели, которая формирует последовательности объединения данных в группы. Также наглядно демонстрируются меры расходимости между соответствующими значениями.

На основании табличных данных (табл.1) и рассчитанной функции плотности нормального закона распределения (формула 4) строится график (рис.3):

$$f(\overline{P}_{m,n}) = \frac{1}{\sqrt{2\cdot\pi}\cdot\delta_{ci}}\cdot e^{-\frac{(\overline{P}_{m,n}-P_{ci})^2}{2\cdot\delta_{ci}^2}} \tag{4}$$

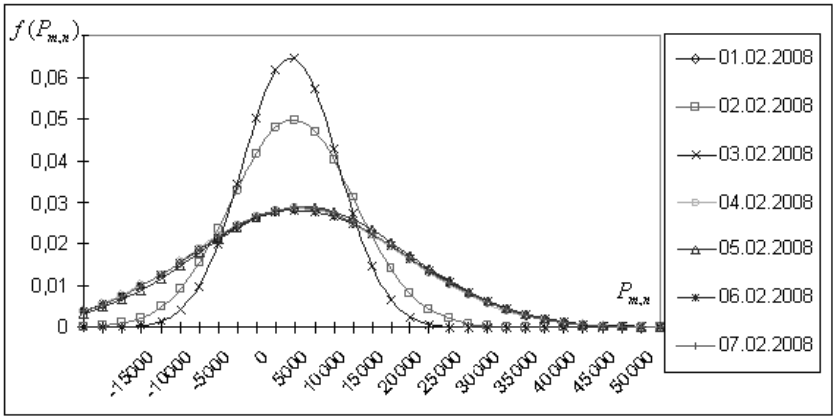

Рис. 3. Плотности нормального закона распределения

Таким образом, можно сделать вывод, что выборка данных носит случайный характер, а закон распределения – нормальный. Следовательно, рассчитанные величины и спроектированные характеристики можно принимать в качестве базисных данных для последующих методов [6].

Кластеризация на основании максимумов плотности вероятности

Кластеризация, которая проводится путем сравнения максимумов плотности вероятности, является одним из наиболее используемых классических методов. Так как величина максимума плотности вероятности $f(\overline{P}_{m,n})_{\max}$ характеризует вероятностное количество значений, сконцентрированных в середине интервала, то чем больше значение $f(\overline{P}_{m,n})_{\max}$, тем равномернее распределение численных значений исходных величин. Применительно к электроэнергетике равномерный график потребления свидетельствует о том, что нагрузка является равномерной без провалов и пиков. Данный случай в основном соответствует выходным дням, когда потребление электроэнергии осуществляется оборудованием, не задействованным в основных технологических процессах.

Соответственно максимальное значение плотности вероятности составляет:

$$f(\overline{P}_{m,n})_{\max} = \frac{1}{\sqrt{2\cdot\pi}\cdot\delta_{ci}}. \tag{5}$$

Классическим способом кластеризации графиков плотности вероятности является сравнение значений $f(\overline{P}_{m,n})_{\max}$ определяемого ряда с заданными диапазонами. Предел устанавливается на границе определения максимумов $f(\overline{P}_{m,n})_{\max}$ для рабочих и выходных дней:

$$f(\overline{P}_{m,n})_{\max} = \frac{f(\overline{P}_{m,n})_{\max p} + f(\overline{P}_{m,n})_{mixв}}{2}, \tag{6}$$

где $f(\overline{P}_{m,n})_{\max p}$ - максимальное значение плотности вероятности, соответствующее максимальному значению для выборки первой группы;

$f(\overline{P}_{m,n})_{mixв}$ - максимальное значение плотности вероятности, соответствующее минимальному значению для выборки второй группы.

Таким образом, устанавливается порог распределения, а в дальнейшем и кластеризации, либо соответствие заданному периоду. На рис.4 изображены установленные пределы в виде диапазонов:

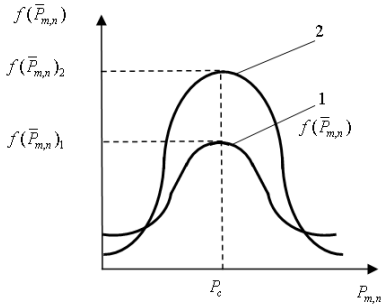

Рис.4. Плотность вероятности нормального закона распределения
1 - для рабочих дней;
2 - для выходных дней

1 $f(\overline{P}_{m,n})_{\max} \in \left[0; \overline{f}(\overline{P}_{m,n})_{\max}\right]$ - рассматриваемая выборка соответствует рабочему дню;

2 $f(\overline{P}_{m,n})_{\max} \in \left[\overline{f}(\overline{P}_{m,n})_{\max}; \infty\right]$ - выборка соответствует выходному дню.

На основании существующих данных:

$$f(\overline{P}_{m,n})_{\max} = \frac{0,065 + 0,027}{2} = 0,046 .$$

Таким образом,

выборка рабочего дня: $f(\overline{P}_{m,n})_{\max} \in [0;\ 0,046]$;

выборка выходного дня: $f(\overline{P}_{m,n})_{\max} \in [0,046;\ \infty]$.

Границы кластеризации определены и применены к базисной выборке(рис.3,табл.1). Полученные результаты представлены на рис.5.

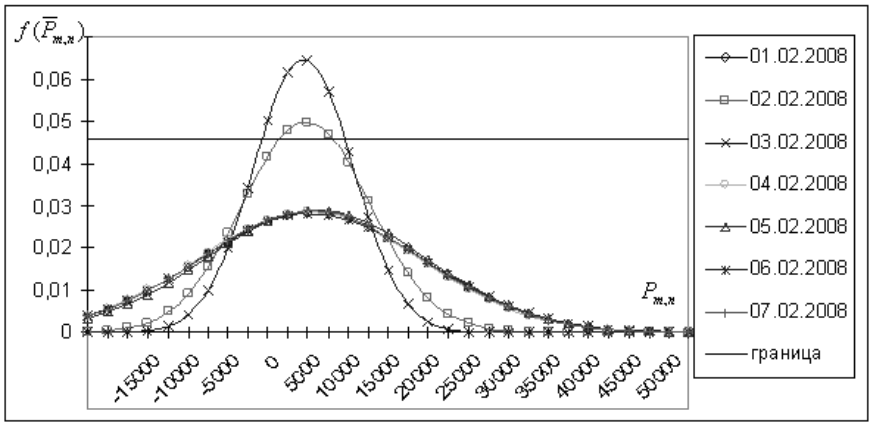

Рис.5. Определение пределов кластеризации плотности вероятности нормального закона распределения

В результате кластеризации установлены две группы:

- рабочий день;
- выходной день.

Проверка данного способа кластеризации графиков электрических нагрузок, проведенная путем ввода контрольной выборки, свидетельствует о его соответствии заданным пределам (рис.6).

Кластеризация графиков электрических нагрузок по касательной к величине плотности вероятности нормального закона распределения

Данный метод основан на том, что к кривой $f(\overline{P}_{m,n})$ проводится касательная до пересечения с прямой, которая проходит из точки $\overline{P}_{ci}$ перпендикулярно оси абсцисс (рис.7). Таким образом устанавливаются диапазоны соответствующих классов $(f(\overline{P}_{m,n})_1, f(\overline{P}_{m,n})_2 ...)$.

При использовании реальных графиков электрических нагрузок данный метод позволил установить два предела (рис.7).

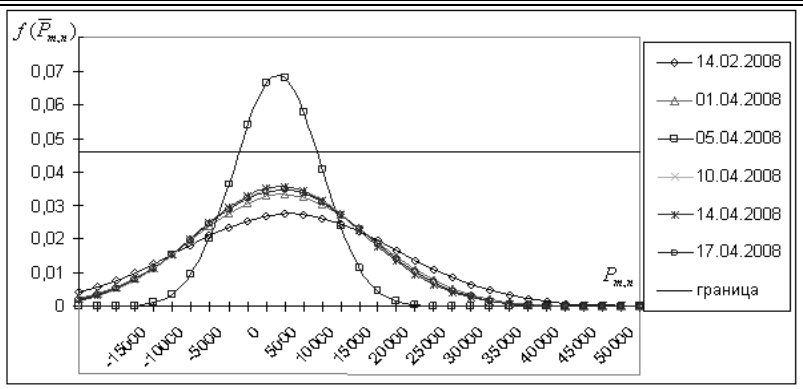

Рис. 6. Проверка выполнения условий кластеризации плотности вероятности нормального закона распределения на основании реальных данных

Для нашего случая эти пределы составляют: для прямой 1 - $f(\overline{P}_{m,n})_1$ =0,038 и прямой 2 - $f(\overline{P}_{m,n})_2$ =0,065. Следует отметить, что при проведении касательных необходимо учитывать влияние изменения потребления для разных групп.

Качество кластеризации с помощью контрольной выборки отмечается высокой точностью (рис.8).

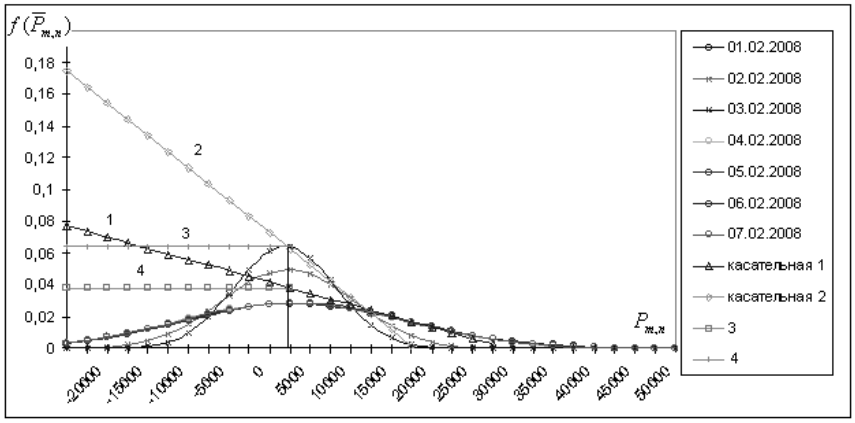

Рис. 7. Установление пределов кластеризации по плотности вероятности нормального закона распределения по касательным

касательная 1 - для рабочих дней; касательная 2 - для выходных дней;

3 – пересечение P_c и касательной 1; 4 - пересечение P_c и касательной 2

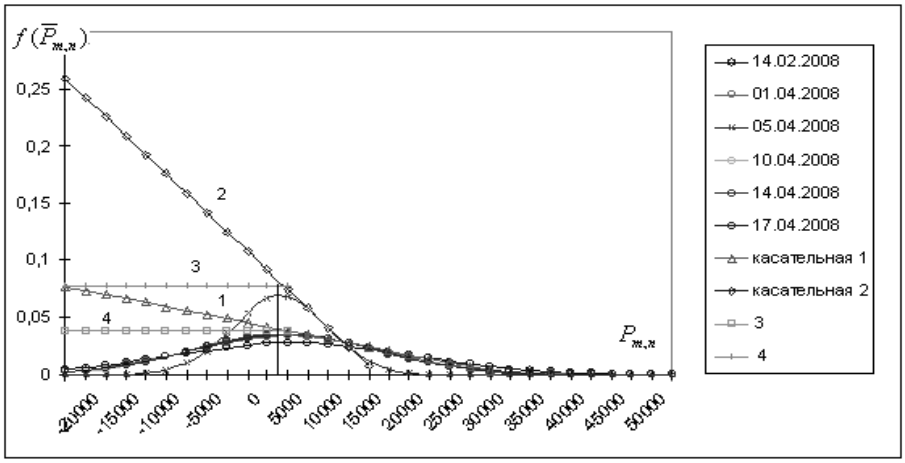

Рис. 8. Проверка выполнения условий кластеризации по касательным к плотности вероятности нормального закона распределения

Выводы

1. Вектор признаков, сформированный на значениях усредненных мощностей вейвлет - коэффициентов по каждой из субполос, является достаточно информативным и может быть использован для кластеризации информационных сигналов.
2. Кластеризация на основе определения максимума плотности вероятности коэффициентов $\overline{P}_{m,n}$ показывает необходимость задания пределов для обеспечения быстрой сходимости ряда.
3. Кластеризация на основе графиков плотности вероятности по касательным также показывает быструю сходимость и восстановление степени отклонения значений от среднего значения в исходном варианте.

Литература

1. Добеши И. Десять лекций по вейвлетам. Москва. - Ижевск: РХД, 2002.
2. Волошко А.В. Метод формирования признаков классификации графиков электрических нагрузок на основе вейвлет-преобразования. // ПРОМЕЛЕКТРО, 2009, № 1. С. 39 - 43.
3. Дворников С.В., Сауков А.М. Метод распознавания радиосигналов на основе вейвлет – пакетов // Научное приборостроение, 2004, т. 14, № 1, С. 57 - 65.
4. Праховник А.В. Прогнозирование предсказания суточных графиков нагрузок на основе адаптивных методов.- Делфт, Нидерланды, 1976, 66 с. (англ.).
5. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001. – 752 с.
6. Бэнн Д.В., Фармер Е.Д. Сравнительные модели прогнозирования электрической нагрузки: Пер. с англ. М.: Энергоатомиздат, 1987. – 200 с.

УДК 621.928.37