

Модели на основе случайных лесов в задаче мониторинга состояний электроэнергетических систем

Жуков А.В.,
Сидоров Д.Н., Томин Н.В., Курбацкий В.Г.,
Панасецкий Д.А., Спиряев В.А.

ИСЭМ СО РАН ИМЭИ ИГУ



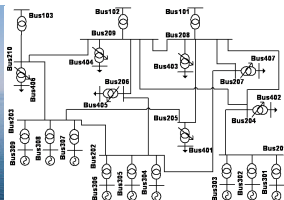
zhukovalex13@gmail.com

ММРО-17
Светлогорск, 2015

План презентации

- 1 **Задачи электроэнергетики**
 - Введение
 - Постановка задачи идентификации состояния
- 2 **Методы решения**
 - Краткий обзор существующих методов
 - Предлагаемые методы
 - Моделирование состояний ЭЭС
- 3 **Тестирование моделей**
 - Тестирование на модельной схеме
- 4 **Результаты и выводы**

Задачи электроэнергетики



Превентивный мониторинг - процедура идентификации и краткосрочного прогнозирования состояния электроэнергосистемы для обнаружения неполадок в системе на раннем этапе с целью их предотвращения.

Цель работы

Цель

Создание системы превентивного мониторинга состояния ЭЭС

Задачи

- анализ существующих мат.методов решения данной задачи и выбор подходящего;
- разработка системы на основе выбранной модели;
- разработка системы моделирования и тестирования;

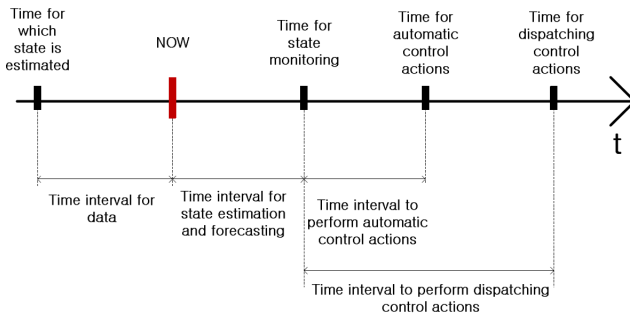
Мотивация

Отсутствие системного подхода к проблеме, необходимость в создании среды прототипирования систем превентивного мониторинга ЭЭС.

Современные электроэнергосистемы (ЭЭС)

- состоят из многих подсистем элементов
- некоторые характеристики (**параметры состояния**) возможно получать в реальном времени
- на нижнем уровне управляется локальной автоматикой
- на верхнем уровне управляется диспетчером

Временные ограничения



Задачи идентификации состояния ЭЭС

Задача идентификации могла бы быть решена с помощью расчёта так называемого **индекса надёжности (Security index) SI** - непрерывная величина, зависит от конфигурации ЭЭС и текущих параметров состояния.

Почему бы не вычислять показатель состояния напрямую?

- вычислительно сложно,
- необходимо обрабатывать пропуски в данных.

Также иногда требуется получить состояние системы в виде:

$$S = (normal, alarm, emergency1, emergency2)$$

Поэтому считается целесообразным использовать модели классификации.

Краткий обзор работ других авторов

Ряд работ посвящённых мониторингу ЭЭС использует **RBFNN, SVM, SOM, Decision Tree (C4.5, J48,...)** для классификации состояний.

Однако всё чаще используются ансамбли деревьев решений:

RSM (Метод случайных подпространств) [Diao et al., 2009]

Boosting [He et al., 2013b, He et al., 2013a]

Random forest [Kamwa et al., 2010, Khoshkhoo and Shahrtash, 2012]

Ошибка не более 1%, но на простых тестовых схемах при простейших сценариях. В реальности всё гораздо сложнее!

Дополнительные ограничения

Какие сложности существуют?

- Невозможность тестирования на реальной системе (доступно только моделирование);
- большое количество признаков;
- жесткие ограничения по времени работы (реальное время);
- нужна обработка пропусков в данных (на этапе классификации);
- число признаков может меняться.

Идея случайных лесов

Случайный лес (Random Forest) - композиционный алгоритм машинного обучения, который использует ансамбль рандомизированных решающих деревьев для решения задач восстановления регрессии или классификации. Он сочетает в себе два фундаментальных принципа:

- Бэггинг** Bootstrap aggregating [L.Brieman, 1996]
Генерация случайных подвыборок (с возвращением) для обучения каждого алгоритма в композиции.
- RSM** Random Subspace Method, [Т.К. Но, 1998]; [Amit and Geman, 1997]
Выбор случайного подмножества признаков для обучения и работы алгоритма.

Формально

Пусть $L = (X_i, Y_i)$, $X \in \mathbb{R}^M$, $Y \in \mathcal{Y}_{i=1}^N$, где X_i – признаки и Y – метки класса $\mathcal{Y} \in 1, 2, \dots, c$, $c \geq 2$.

Изначальная реализация случайного леса равномерно и независимо выбирая элементы из L строит для обучения каждого k -го дерева T_k^* новую выборку L_k^* . Этот процесс повторяется заданное количество раз K . Итоговое значение выбирается путём **голосования по большинству**. Так для конкретного $X = x$, решение в случае классификации выглядит как

$$\hat{H}(x) = \operatorname{argmax}_{y \in \mathcal{Y}} \sum_{k=1}^K I(\hat{h}_k(x) = y)$$

для классификации, где $\hat{h}_k(x)$ решающее правило k -го решающего дерева ансамбля, а I – индикаторная функция.

Почему Random Forests?

- низкая ошибка сопоставимая с SVM, GBM;
- возможность распараллеливания;
- встроенный метод оценивания ошибки обобщения (OOB-test) и оценки значимости признаков;
- имеется возможность работы с частичными пропусками данных;



Диаграмма обработки данных



Создание начальной обучающей выборки

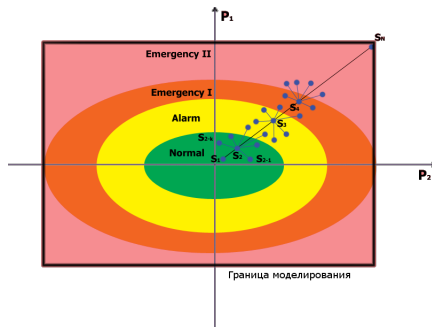
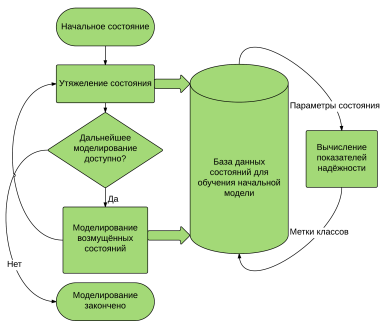


Figure: Схема создания базы состояний ЭЭС

Модель ЭЭС

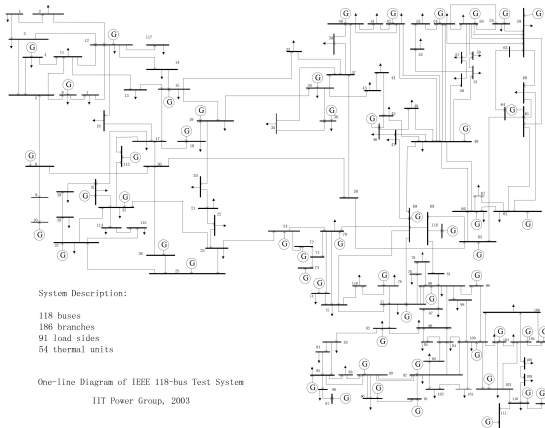


Figure: Схема модельной ЭЭС IEEE 118 линий

Матрица несоответствий

	alarm	emergency1	emergency2	normal	class.error
alarm	947.00	1.00	0.00	0.00	0.00
emergency1	4.00	279.00	5.00	0.00	0.03
emergency2	0.00	4.00	252.00	0.00	0.02
normal	2.00	0.00	0.00	225.00	0.01

Table: Матрица несоответствий для схемы IEEE 118

Кривые обучения для RF на IEEE 118 buses

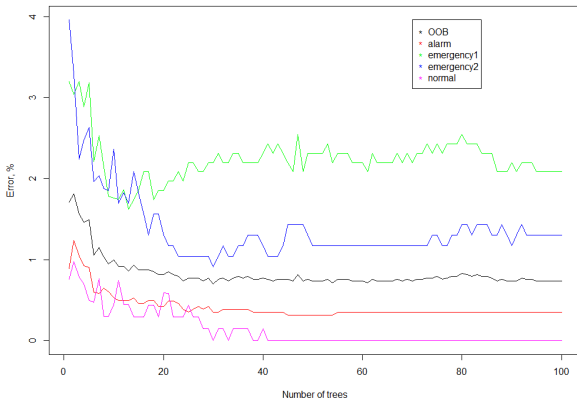


Figure: Кривые ошибки при обучении и тестировании

Значимость переменных

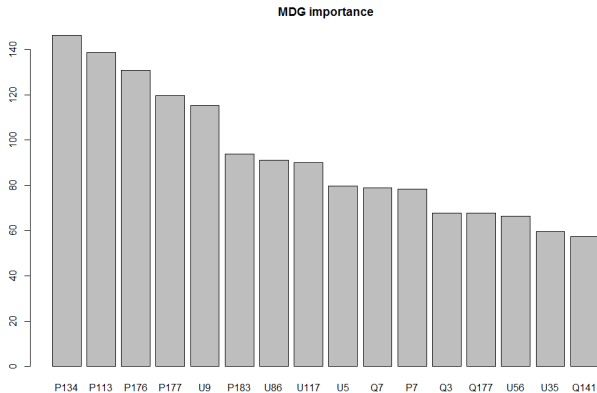


Figure: Относительная значимость признаков (Mean decrease Gini importance)

Результаты работы

Результаты

- разработана система прототипирования моделей превентивного мониторинга
- разработан метод периодического обновления случайного леса

Дальнейшие планы

- адаптация Online-моделей для данной задачи

Вопросы



e-mail: zhukovalex13@gmail.com

Вычислительная сложность случайных лесов

Алгоритм	Лучший	Худший	В среднем
CART	$\mathcal{O}(pN \log^2 N)$	$\mathcal{O}(pN^2 \log N)$	$\mathcal{O}(pN \log^2 N)$
Bagging	$\mathcal{O}(Mp\tilde{N} \log^2 \tilde{N})$	$\mathcal{O}(Mp\tilde{N}^2 \log \tilde{N})$	$\mathcal{O}(Mp\tilde{N} \log^2 \tilde{N})$
RF	$\mathcal{O}(MK\tilde{N} \log^2 \tilde{N})$	$\mathcal{O}(MK\tilde{N}^2 \log \tilde{N})$	$\mathcal{O}(MK\tilde{N} \log^2 \tilde{N})$
ETs	$\mathcal{O}(MKN \log N)$	$\mathcal{O}(MKN^2)$	$\mathcal{O}(MKN \log N)$
PERT	$\mathcal{O}(MN \log N)$	$\mathcal{O}(MN^2)$	$\mathcal{O}(MN \log N)$

Table: Асимптотические оценки вычислительной сложности

M - количество деревьев,

N - длина выборки,




p - количество входных переменных,

K - количество переменных выбираемых для поиска разбиения,




\tilde{N} - количество примеров используемых в бутстреппе ($\approx 63\%$ от N)

[Loupre, 2014]

References I

-  Amit, Y. and Geman, D. (1997).
Shape quantization and recognition with randomized trees.
Neural computation, 9(7):1545–1588.
-  Diao, R., Sun, K., Vittal, V., O’Keefe, R. J., Richardson, M. R.,
Bhatt, N., Stradford, D., and Sarawgi, S. K. (2009).
Decision tree-based online voltage security assessment using pmu
measurements.
Power Systems, IEEE Transactions on, 24(2):832–839.
-  He, M., Vittal, V., and Zhang, J. (2013a).
Online dynamic security assessment with missing pmu
measurements: A data mining approach.
Power Systems, IEEE Transactions on, 28(2):1969–1977.

References II

-  He, M., Zhang, J., and Vittal, V. (2013b). Robust online dynamic security assessment using adaptive ensemble decision-tree learning. *Power Systems, IEEE Transactions on*, 28(4):4089–4098.
-  Kamwa, I., Samantaray, S., and Joos, G. (2010). Catastrophe predictors from ensemble decision-tree learning of wide-area severity indices. *Smart Grid, IEEE Transactions on*, 1(2):144–158.
-  Khoshkhou, H. and Shahrtash, S. (2012). On-line dynamic voltage instability prediction based on decision tree supported by a wide-area measurement system. *IET Generation, Transmission & Distribution*, 6(11):1143–1152.

References III



Louppe, G. (2014).

Understanding random forests: From theory to practice.
arXiv preprint arXiv:1407.7502.